L'attenzione come operatore di selezione del gradiente

Quarta monografia del ciclo *Architettura della mente*

Nota conclusiva sul ciclo

Questa è la quarta e ultima monografia del ciclo *Architettura della mente*. Le quattro monografie condividono un impianto tecnico comune — la mente come sistema che si auto-rappresenta in modo irriducibilmente parziale — e ciascuna ne sviluppa un aspetto attraverso un apparato formale distinto.

La prima — *Il malinteso come struttura cognitiva* — ha mostrato che la comunicazione tra sistemi con rappresentazioni diverse non converge necessariamente al significato condiviso: il malinteso è il punto fuori equilibrio del gioco comunicativo, la cui struttura di Nash è il fondamento del significato intersoggettivo anziché la sua degradazione. L'apparato è venuto dalla teoria dei giochi della comunicazione e dalla semantica formale.

La seconda — *Pensare senza parole* — ha costruito il substrato subpropositivo come sistema di rappresentazione distribuita in $\mathbb{R}^n$, di cui il linguaggio è un codec lossy. La relazione substrato-linguaggio non è unidirezionale: Φ^{-1} è un operatore che rimodella la geometria di $\mathbb{R}^n$ attraverso l'uso ripetuto, e il sistema dinamico accoppiato substrato-linguaggio ha come attrattore la struttura cognitiva stabile dell'esperto. L'apparato è venuto dalla teoria delle rappresentazioni distribuite e dalla geometria degli spazi di attivazione.

La terza — *Metacognizione: il modello del modello* — ha trattato la metacognizione come compressione di secondo ordine soggetta ai teoremi di impossibilità per l'auto-conoscenza formale. Gödel, Tarski, Löb, Rice, Chaitin e il no-free-lunch sono istanze dello schema diagonale di Lawvere: qualsiasi tentativo di auto-rappresentazione completa produce il punto fisso impossibile. La logica GL è la geometria propria dell'auto-conoscenza coerente; gli ordinali proof-teorici ε_K misurano la portata metacognitiva; il prior di Solomonoff è il limite infattibile dell'ottimalità bayesiana. L'apparato è venuto dalla teoria bayesiana gerarchica e dai teoremi di impossibilità.

La quarta monografia — il testo che segue — tratta l'attenzione non come risorsa scarsa da allocare ottimamente ma come selettore di quale prior viene aggiornato: il meccanismo attraverso cui il flusso dell'esperienza viene convertito in aggiornamento

selettivo delle credenze. La tesi centrale è che chi controlla il processo di selezione del gradiente — chi determina quale porzione del flusso informazionale riceve attenzione e quale viene soppressa — controlla l'evoluzione delle credenze, tanto individuali quanto collettive. Le conseguenze di questa tesi attraversano la psicologia cognitiva, la teoria dell'informazione, e l'economia politica formale dei sistemi cognitivi collettivi.

Il filo diagonale che attraversa il ciclo trova qui la propria chiusura: l'auto-attenzione — il meccanismo che seleziona i prior da aggiornare è esso stesso soggetto all'aggiornamento che seleziona — produce la struttura di punto fisso che è la firma formale della coscienza come attrattore dinamico e della manipolazione come deformazione del bacino di attrazione.

Proemio: La geometria dell'aggiornamento selettivo

C'è una domanda che la teoria standard dell'attenzione non pone perché presuppone di avere già risposto: perché l'attenzione è preziosa? La risposta corrente — perché le risorse cognitive sono limitate e l'attenzione è il meccanismo di allocazione di quelle risorse — è vera ma superficiale. Tratta l'attenzione come un'economia di scarsità: tanti bit disponibili, tanti ne vengono elaborati, l'attenzione decide quali. In questa cornice, il problema dell'attenzione è un problema di priorità: quale informazione è più urgente, quale segnale ha il rapporto segnale-rumore migliore, quale compito ha l'utilità marginale più alta.

Questa concezione ha radici filosofiche solide — William James la formula con precisione nel 1890, e la ricerca sperimentale del Novecento ne ha elaborato le implicazioni con raffinatezza crescente — ma manca la struttura più profonda del fenomeno. L'attenzione non è soltanto un filtro che decide quale informazione entra nel sistema: è il meccanismo che determina *in quale direzione* il sistema si aggiorna. Dati gli stessi stimoli fisici, direzioni diverse dell'attenzione producono aggiornamenti diversi delle credenze — non soltanto aggiornamenti di più o di meno, ma aggiornamenti qualitativamente distinti, che modificano dimensioni diverse dello spazio delle credenze e portano il sistema verso regioni diverse del suo spazio degli stati.

La formalizzazione precisa di questa struttura richiede il linguaggio dell'ottimizzazione stocastica. In quell'ambiente, l'aggiornamento delle credenze è un passo di discesa del gradiente su una funzione di perdita nello spazio delle distribuzioni di probabilità — il *gradiente bayesiano*, la direzione che minimizza la divergenza KL tra la distribuzione corrente e la distribuzione ottimale dato l'input osservato. L'attenzione, in questa formalizzazione, è l'operatore che *seleziona* quale componente del gradiente viene effettivamente applicata: quale dimensione dello spazio delle credenze viene aggiornata in risposta a quale porzione dell'input.

Questa selezione del gradiente non è una questione tecnica di implementazione: è una questione di potere epistemico. Chi controlla quale gradiente viene applicato controlla la traiettoria del sistema nello spazio delle credenze. In un sistema cognitivo individuale, il controllo dell'attenzione è il meccanismo attraverso cui certi prior vengono sistematicamente rinforzati e altri sistematicamente ignorati — la struttura formale della formazione dell'identità cognitiva, del bias cognitivo, e della trasformazione terapeutica. In un sistema cognitivo collettivo — una società, un'organizzazione, un mercato dell'informazione — il controllo dell'attenzione è il meccanismo di potere epistémico fondamentale: il punto di controllo attraverso cui si determina quale distribuzione di credenze una collettività di agenti converge.

L'argomento si sviluppa in cinque movimenti. Il primo costruisce la formalizzazione dell'attenzione come operatore di selezione del gradiente bayesiano, precisando la struttura matematica dell'aggiornamento selettivo. Il secondo analizza la geometria dell'aggiornamento nello spazio delle distribuzioni di probabilità attraverso il mirror descent e il gradiente naturale, mostrando come la curvatura dello spazio riemanniano delle distribuzioni modifica l'effetto dell'attenzione. Il terzo tratta il controllo dell'attenzione come controllo dell'evoluzione delle credenze, con la formalizzazione della propaganda, della persuasione e della manipolazione come operatori sullo spazio dei gradienti. Il quarto estende l'analisi ai sistemi collettivi attraverso l'economia politica formale: i mercati dell'attenzione come meccanismi di aggregazione delle traiettorie cognitive, con i teoremi di impossibilità per l'aggregazione equa. Il quinto chiude il ciclo con l'auto-attenzione come struttura di punto fisso diagonale: il momento in cui l'attenzione si rivolge su se stessa e produce la struttura che la terza monografia ha identificato come il limite di ogni auto-conoscenza formale.

Capitolo I — L'attenzione come operatore di selezione del gradiente bayesiano

I.1 Il gradiente bayesiano: definizione e struttura

La teoria bayesiana della cognizione tratta le credenze di un agente come distribuzioni di probabilità su uno spazio di ipotesi, aggiornate in risposta a nuove osservazioni secondo la regola di Bayes. Questo framework è standard — è la struttura del Capitolo I della terza monografia — ma la sua formulazione come problema di ottimizzazione non è stata esplicitata con la precisione necessaria per il presente scopo.

Sia P lo spazio delle distribuzioni di probabilità su uno spazio di ipotesi H. Sia $\pi_t \in P$ la credenza dell'agente al tempo t. Sia o_{t+1} un'osservazione. L'aggiornamento bayesiano è:

$$\pi_{t+1}(h) = P(o_{t+1} \mid h) \cdot \pi_t(h) / Z_{t+1}$$

dove $Z_{t+1} = \int_H P(o_{t+1} \mid h)\, d\pi_t(h)$ è la normalizzazione. Questa formula è il risultato di un problema di ottimizzazione: π_{t+1} è il minimizzatore della divergenza KL dal prior π_t soggetto al vincolo che la verosimiglianza attesa dell'osservazione sia massimizzata:

$\pi_{t+1} = \operatorname{argmin}_{\pi \in P} D_{KL}(\pi \,\|\, \pi_t)$ soggetto a $\mathbb{E}_\pi[\log P(o_{t+1} \mid h)] =$ massimo

Il gradiente di questa funzione obiettivo nello spazio P — il *gradiente bayesiano* — ha la struttura:

$$\nabla_P \mathcal{F}(\pi ; o) = \log P(o \mid \cdot) - \mathbb{E}_\pi[\log P(o \mid h)]$$

il log-rapporto di verosimiglianza centrato: la differenza tra la log-verosimiglianza dell'osservazione ipotesi per ipotesi e la sua media sotto la distribuzione corrente. Il gradiente punta nella direzione delle ipotesi che spiegherebbero meglio l'osservazione rispetto alla media attuale.

Il passo di aggiornamento bayesiano è il passo di discesa del gradiente nel senso della geometria naturale dello spazio P — come si elaborerà nel Capitolo II. Il punto cruciale è che il gradiente bayesiano è un oggetto in P: è una funzione sull'insieme delle ipotesi, non un scalare. Ogni dimensione del gradiente corrisponde all'aggiornamento di una porzione dello spazio delle ipotesi H.

I.2 L'attenzione come maschera sul gradiente

Si formalizza ora l'attenzione come operatore che agisce sul gradiente bayesiano. Sia $A_t : H \to [0,1]$ una *maschera di attenzione* — una funzione che a ogni ipotesi $h \in H$ assegna un peso di attenzione $A_t(h) \in [0,1]$. Il *gradiente attenzionato* è:

$$\nabla_P^A \mathcal{F}(\pi_t ; o_{t+1})(h) = A_t(h) \cdot [\log P(o_{t+1} \mid h) - \mathbb{E}_\pi[\log P(o \mid h)]]$$

il gradiente bayesiano moltiplicato componente per componente dalla maschera di attenzione. L'aggiornamento attenzionato è il passo di discesa del gradiente attenzionato anziché del gradiente completo:

$$\pi_{t+1}(h) \propto \pi_t(h) \cdot \exp(\eta \cdot A_t(h) \cdot \log P(o_{t+1} \mid h))$$

dove $\eta > 0$ è il tasso di apprendimento. Per $A_t \equiv 1$ (attenzione piena su tutto lo spazio delle ipotesi) si recupera l'aggiornamento bayesiano standard. Per A_t a supporto ristretto — attenzione concentrata su una sottoregione $H_A \subset H$ — si ottiene un aggiornamento parziale che modifica la distribuzione sulle ipotesi in H_A senza toccare quelle in $H \setminus H_A$.

La maschera di attenzione A_t non è esogena: dipende dall'osservazione o_{t+1}, dalla credenza corrente π_t, e dalla struttura del sistema di attenzione. Un agente razionale con risorse limitate costruirebbe A_t come soluzione di un problema di ottimizzazione: data una capacità totale di attenzione $\int_H A_t(h)\, dh \leq C$, quale A_t massimizza l'utilità attesa dell'aggiornamento?

I.3 L'ottimizzazione della maschera di attenzione

Il problema di ottimizzazione dell'attenzione è: dati π_t, o_{t+1}, e la capacità C, trovare la maschera ottimale A_t*:

$$A_t^* = \operatorname{argmax}_{A : \int A\, dh \leq C} \mathcal{V}(\pi_{t+1}^A)$$

dove $\mathcal{V}(\pi)$ è il valore dell'agente con credenza π — tipicamente la somma scontata delle utilità attese future sotto quella distribuzione di credenze — e π_{t+1}^A è la credenza aggiornata mediante il gradiente attenzionato con maschera A.

Il valore della credenza $\mathcal{V}(\pi)$ è una funzione concava su P per agenti con funzione di utilità concava sull'insieme delle azioni. La variazione di valore indotta dall'aggiornamento con maschera A è:

$$\Delta\mathcal{V}(A) = \mathcal{V}(\pi_{t+1}^A) - \mathcal{V}(\pi_t) \approx \langle \nabla_P \mathcal{V}(\pi_t),\ \nabla_P^A \mathcal{F}(\pi_t ; o) \rangle$$

il prodotto interno tra il gradiente del valore e il gradiente bayesiano attenzionato. La maschera ottimale risolve il problema:

$$A_t^* = \operatorname{argmax}_{A : \int A \leq C} \int_H A(h) \cdot [\nabla_P \mathcal{V}(\pi_t)(h)] \cdot [\log P(o \mid h) - \mathbb{E}_\pi]\, dh$$

Questo è un problema di programmazione lineare continua la cui soluzione — per un vincolo di budget sulla capacità totale di attenzione — è la maschera bang-bang: A_t*(h) = 1 per le ipotesi h in cui il prodotto [valore marginale] · [gradiente bayesiano] è sopra una soglia, e A_t*(h) = 0 altrove.

La struttura bang-bang dell'attenzione ottimale corrisponde precisamente alla fenomenologia sperimentale dell'attenzione selettiva: l'attenzione non si distribuisce uniformemente ma si concentra su ciò che è più informativo *per l'agente* — non soltanto su ciò che ha il maggiore contenuto informazionale assoluto. La soglia di selezione dipende sia dalla struttura del gradiente bayesiano (informatività dell'osservazione per ogni ipotesi) sia dal gradiente del valore (rilevanza di ogni ipotesi per le decisioni future). Un'ipotesi altamente informata ma irrilevante per le azioni future non merita attenzione; un'ipotesi poco informata ma cruciale per una decisione imminente la merita.

I.4 Il costo dell'attenzione selettiva: perdita di calibrazione

L'aggiornamento attenzionato con maschera $A_t \neq 1$ produce una distribuzione di credenze π_{t+1}^A che differisce dall'aggiornamento bayesiano completo π_{t+1}^{Bayes}. La distanza tra le due:

$$D_{KL}(\pi_{t+1}^{Bayes} \| \pi_{t+1}^A) = \mathbb{E}_{Bayes}[\log P(o|h)] - \mathbb{E}_A[\log P(o|h)]$$

misura la perdita di calibrazione indotta dall'attenzione selettiva — la distanza tra la distribuzione ottimale e quella prodotta dall'attenzione limitata. Questa perdita è il costo dell'attenzione selettiva: il prezzo pagato in accuratezza delle credenze per la riduzione del costo computazionale dell'elaborazione.

La perdita di calibrazione ha una struttura importante: è concentrata sulle ipotesi in $H \setminus H_A$ — quelle che ricevono $A_t(h) = 0$. Per quelle ipotesi, l'aggiornamento è nullo: la distribuzione π_{t+1}^A su $H \setminus H_A$ è identica a π_t su $H \setminus H_A$. L'osservazione o_{t+1} non ha modificato le credenze sulle ipotesi non attenzionate, anche se la corretta applicazione della regola di Bayes le avrebbe modificate.

Questo produce un bias sistematico nelle credenze dell'agente con attenzione limitata: le ipotesi che sistematicamente sfuggono all'attenzione vengono sistematicamente sotto- o sovra-stimate rispetto alla distribuzione bayesiana ottimale. Se l'attenzione è costantemente concentrata su un sottoinsieme delle ipotesi $H_A \subset H$, il sistema converge a una distribuzione di credenze π^* che è calibrata su H_A ma distorta su $H \setminus H_A$. La distorsione si accumula nel tempo:

$$D_{KL}(\pi_T^{Bayes} \| \pi_T^A) \leq T \cdot \max_t D_{KL}(\pi_{t+1}^{Bayes} \| \pi_{t+1}^A)$$

la divergenza totale è bounded linearmente nel numero di passi di aggiornamento per la massima perdita per passo. Per agenti con attenzione sistematicamente distorta verso H_A, la traiettoria delle credenze diverge irreversibilmente da quella bayesiana ottimale: la storia delle distorsioni attenzionali determina la struttura delle credenze finali.

I.5 Attenzione e struttura dello spazio delle ipotesi: la topologia dell'aggiornamento

La maschera di attenzione A_t induce una partizione dello spazio delle ipotesi in due insiemi: H_A (attenzionato) e $H \setminus H_A$ (non attenzionato). Questa partizione non è arbitraria: è determinata dalla struttura del sistema di attenzione e dall'input.

Si osserva che la maschera di attenzione definisce una *topologia* sullo spazio delle ipotesi: due ipotesi $h_1, h_2 \in H$ sono *attenzionalmente vicine* se $A_t(h_1)$ e $A_t(h_2)$ sono simili per tutte le osservazioni tipiche $o_{t+1} \sim P(o \mid \pi_t)$. La topologia attenzionale è la struttura che determina quali distinzioni tra ipotesi vengono percepite come rilevanti e quali vengono trascurate.

La topologia attenzionale ha una connessione precisa con la geometria dello spazio di attivazione $\mathbb{R}^n$ della seconda monografia. La maschera di attenzione nello spazio delle ipotesi corrisponde, nello spazio di attivazione, alla curvatura locale della varietà cognitiva M: le regioni di M con alta curvatura corrispondono a sottoinsieme di ipotesi che il sistema distingue finemente; le regioni piatte corrispondono a ipotesi indistinguibili. L'attenzione orienta la curvatura della varietà cognitiva verso le regioni ipotetico-spaziali rilevanti per l'agente, con la conseguenza che il sistema di credenze dell'agente sviluppa progressivamente la struttura di un esperto nel dominio su cui l'attenzione è stata sistematicamente concentrata.

Questa connessione formalizza il meccanismo attraverso cui l'expertise percettiva emerge dall'attenzione sostenuta: la concentrazione dell'attenzione su un sottoinsieme delle ipotesi H_A orienta sia l'aggiornamento delle credenze (attraverso il gradiente attenzionato) sia la curvatura dello spazio di attivazione nella regione corrispondente (attraverso il meccanismo f_lang della seconda monografia). Expertise e attenzione sono co-evolutivi: l'expertise aumenta la risoluzione dell'attenzione nelle dimensioni rilevanti, e l'attenzione sostenuta costruisce l'expertise modificando la geometria del substrato.

I.6 L'attenzione come operatore in forma di semigruppe

Si definisce formalmente la struttura algebrica dell'attenzione come operatore sullo spazio delle credenze. La composizione di due passi di aggiornamento attenzionato è:

$$\pi_{t+2} = Upd^{A_{t+1}}(Upd^{A_t}(\pi_t, o_{t+1}), o_{t+2})$$

dove $Upd^A(\pi, o)(h) \propto \pi(h) \cdot \exp(\eta \cdot A(h) \cdot \log P(o|h))$. Si definisce l'*operatore di attenzione* come il generatore infinitesimale del flusso di aggiornamento: nella limite di passi infinitesimali $\delta t \to 0$, il flusso è:

$$d\pi_t/dt = \mathscr{A}_t[\pi_t] = -\pi_t \cdot A_t \cdot (\log P(O_t \mid \cdot) - \mathbb{E}_{\pi_t}[\log P(O_t|\cdot)])$$

dove O_t è il processo di osservazione. L'operatore $\mathscr{A}_t$ è un operatore lineare su P — applicato a una distribuzione π, produce un campo di variazione su H — la cui struttura dipende dalla maschera A_t e dall'input O_t.

L'insieme degli operatori $\{\mathscr{A}_t : t \geq 0\}$ genera un semigruppe di trasformazioni su P:

$$\{\Phi_t = \exp(\int_0^t \mathscr{A}_s \, ds) : t \geq 0\}$$

Il semigruppe dell'attenzione descrive la dinamica completa del sistema di credenze sotto un processo di osservazione con attenzione variabile nel tempo. La struttura del semigruppe — i suoi generatori, i loro commutatori, i punti fissi — è la struttura formale del processo di formazione delle credenze.

Un punto fisso π^* di Φ_t è una distribuzione di credenze stabile: le osservazioni che riceve attenzione non la spostano (o meglio, la spostano in un ciclo stabile). Questa è la struttura formale dell'identità cognitiva: un sistema di credenze che si auto-mantiene attraverso il processo di aggiornamento selettivo. L'analisi dei punti fissi del semigruppe dell'attenzione è il fondamento formale della psicologia dei bias cognitivi — strutture di credenza che si auto-rinforzano attraverso la selezione dell'attenzione verso le ipotesi compatibili — e della trasformazione terapeutica — la modifica della maschera di attenzione che porta a un diverso punto fisso.

Capitolo II — La geometria dell'aggiornamento: mirror descent e gradiente naturale

II.1 Lo spazio delle distribuzioni come varietà riemanniana

Il framework del Capitolo I tratta l'aggiornamento bayesiano come discesa del gradiente nello spazio P delle distribuzioni di probabilità. Ma P non è uno spazio euclideo piatto: è una varietà riemanniana con una struttura geometrica propria che determina la "forma" corretta della discesa del gradiente. Ignorare questa struttura — usare il gradiente euclideo ordinario anziché il gradiente naturale — produce un algoritmo di aggiornamento che converge lentamente e che è sensibile alla parametrizzazione scelta per le distribuzioni.

La geometria naturale dello spazio P è quella definita dalla metrica di Fisher. Per uno spazio di distribuzioni parametrizzate $\{P_\theta : \theta \in \Theta\}$, la *metrica di Fisher* è il tensore:

$$g_{ij}(\theta) = \mathbb{E}_{P_\theta}[\partial_i \log P_\theta(h) \cdot \partial_j \log P_\theta(h)]$$

$$= -\mathbb{E}_{P_\theta}[\partial_i \partial_j \log P_\theta(h)]$$

dove $\partial i = \partial/\partial\theta_i$. *La matrice di informazione di Fisher* $G(\theta) = (g_{ij}(\theta))$ definisce la lunghezza degli spostamenti nello spazio dei parametri: lo spostamento $d\theta$ ha lunghezza $\|d\theta\|_F = \sqrt{(d\theta^T G(\theta)\, d\theta)}$.

La metrica di Fisher ha un significato preciso in termini di divergenza KL: è la curvatura locale della divergenza KL tra distribuzioni vicine:

$$D_{KL}(P_\theta \,\|\, P_{\theta+d\theta}) = \tfrac{1}{2}\, d\theta^T G(\theta)\, d\theta + O(\|d\theta\|^3)$$

Lo spostamento nello spazio dei parametri che produce la massima riduzione della divergenza KL per lunghezza unitaria è il *gradiente naturale*:

$$\tilde{\nabla}_\theta \mathcal{F} = G(\theta)^{-1} \nabla_\theta \mathcal{F}$$

il gradiente ordinario pre-moltiplicato dall'inversa della matrice di Fisher. Il gradiente naturale è il gradiente riparametrizzato che tiene conto della geometria riemanniana dello spazio delle distribuzioni.

II.2 Il mirror descent come generalizzazione dell'aggiornamento bayesiano

Il mirror descent è un algoritmo di ottimizzazione che generalizza la discesa del gradiente a spazi con struttura non euclidea, usando una funzione di distanza di Bregman al posto della distanza euclidea. Nel contesto dell'aggiornamento delle credenze, il mirror descent produce una famiglia di aggiornamenti che include l'aggiornamento bayesiano standard come caso particolare.

Sia $\Psi : P \to \mathbb{R}$ una funzione convessa e differenziabile — il *potenziale di Bregman*. La divergenza di Bregman associata è:

$$D_\Psi(\pi \| \sigma) = \Psi(\pi) - \Psi(\sigma) - \langle \nabla\Psi(\sigma), \pi - \sigma \rangle$$

la differenza tra $\Psi(\pi)$ e la sua approssimazione lineare centrata in σ. Il passo di mirror descent con potenziale Ψ, tasso η, e gradiente $\nabla \mathcal{F}$ è:

$$\pi_{t+1} = \text{argmin}_{\pi \in P} \{ \eta \cdot \langle \nabla \mathcal{F}(\pi_t), \pi \rangle + D_\Psi(\pi \| \pi_t) \}$$

Per $\Psi = H$ (l'entropia negativa, $\Psi(\pi) = \Sigma_h \pi(h) \log \pi(h)$), la divergenza di Bregman è la divergenza KL:

$$D_H(\pi \| \sigma) = D_{KL}(\pi \| \sigma)$$

e il passo di mirror descent diventa:

$$\pi_{t+1}(h) \propto \pi_t(h) \cdot \exp(-\eta \cdot \partial\mathcal{F}/\partial\pi(h))$$

Per $\mathcal{F}(\pi) = -\mathbb{E}_\pi[\log P(o|h)]$ (la log-verosimiglianza negativa), il passo diventa:

$$\pi_{t+1}(h) \propto \pi_t(h) \cdot \exp(\eta \cdot \log P(o|h)) = \pi_t(h) \cdot P(o|h)^\eta$$

Per $\eta = 1$ si ottiene esattamente l'aggiornamento bayesiano. Per $\eta \neq 1$ si ottengono generalizzazioni: $\eta < 1$ produce aggiornamenti "conservativi" che si discostano poco dal prior; $\eta > 1$ produce aggiornamenti "aggressivi" che si discostano molto.

Il mirror descent fornisce la struttura formale che unifica l'aggiornamento bayesiano standard con le sue generalizzazioni nei modelli di bounded rationality: un agente con capacità limitata di elaborazione delle verosimiglianze non applica $\eta = 1$ ma un $\eta_{eff} < 1$ che riflette la compressione dell'informazione imposta dai limiti computazionali.

La maschera di attenzione A_t del Capitolo I produce effettivamente un η_eff(h) = A_t(h) · η che varia per ipotesi: le ipotesi attenzionate ricevono l'aggiornamento pieno, quelle non attenzionate ricevono η_eff = 0.

II.3 Il gradiente naturale attenzionato: la struttura riemanniana dell'attenzione

La combinazione della geometria riemanniana di Fisher con la selezione attenzionale del gradiente produce il *gradiente naturale attenzionato*:

$\nabla^{\sim}$_P^A $\mathcal{F}(\pi ; o)(h) = [G(\pi)^{-1} \cdot \nabla$_P^A $\mathcal{F}(\pi ; o)](h)$

il gradiente attenzionato pre-moltiplicato dall'inversa della matrice di Fisher. Questo è il gradiente che un agente razionale con attenzione selettiva e consapevolezza della geometria riemanniana dello spazio delle credenze dovrebbe seguire.

La matrice di Fisher G(π) ha una struttura che dipende dalla distribuzione corrente π. In regioni di P dove la distribuzione è concentrata (alta confidenza in poche ipotesi), G(π) ha autovalori piccoli nelle direzioni ortogonali al supporto — l'aggiornamento lungo quelle direzioni è amplificato dall'inversione di G. In regioni dove la distribuzione è diffusa (alta incertezza), G(π) ha autovalori uniformi — l'aggiornamento è isotropico.

La conseguenza per l'attenzione è la seguente. La maschera ottimale A_t* del Capitolo I è derivata ignorando la geometria riemanniana. Quando si tiene conto della geometria, la maschera ottimale si modifica: le ipotesi in regioni di P con alta curvatura (dove G ha autovalori grandi) richiedono meno attenzione per produrre lo stesso aggiornamento — la geometria "amplifica" il gradiente in quelle direzioni. Il problema di ottimizzazione dell'attenzione corretto è:

A_t* = argmax_{A : $\int$ A ≤ C} $\int$_H A(h) · [∇_P $\mathcal{V}(\pi_t)$] · G(π_t)$^{-1}$ · [log P(o|h) - $\mathbb{E}$_π] dh

La maschera ottimale tiene conto della geometria: le ipotesi in regioni di alta curvatura sono "economicamente raggiungibili" e richiedono meno attenzione per un dato aggiornamento; quelle in regioni piatte richiedono più attenzione.

II.4 La connessione con la geometria dell'informazione: la triplice struttura duale

La geometria dell'informazione (Amari, 1985) identifica una struttura duale nello spazio delle distribuzioni che è fondamentale per comprendere la geometria dell'aggiornamento attenzionato. Per ogni coppia (π, σ) ∈ P × P si hanno due divergenze:

$D^{+}(\pi \| \sigma) = D_{KL}(\pi \| \sigma)$ (divergenza "in avanti")

$D^{-}(\pi \,\|\, \sigma) = D_{KL}(\sigma \,\|\, \pi)$ (divergenza "all'indietro")

che inducono due connessioni affini sullo spazio P: la connessione e-piatta (esponenziale) e la connessione m-piatta (mista). Le geodetiche e-piatte corrispondono alle traiettorie di aggiornamento bayesiano in cui si minimizza $D_{KL}(P_{target} \,\|\, \pi)$ — il problema del momento (moment matching); le geodetiche m-piatte corrispondono alle traiettorie in cui si minimizza $D_{KL}(\pi \,\|\, P_{target})$ — il problema della proiezione.

L'aggiornamento bayesiano standard è una proiezione m-piatta: π_{t+1} è la proiezione m-piatta di π_t sull'iperpiano $\mathcal{L}_o = \{\pi : \mathbb{E}_\pi[\log P(o|h)] = \text{massimo}\}$. L'aggiornamento attenzionato con maschera A_t è una proiezione m-piatta su un iperpiano modificato dalla maschera:

$\mathcal{L}_o^A = \{\pi : \int_H A(h) \cdot \mathbb{E}_\pi[\log P(o|h)]\, dh = \text{massimo}\}$

La geometria della proiezione m-piatta spiega perché l'aggiornamento bayesiano attenzionato non è commutativo: cambiando l'ordine di due aggiornamenti con maschere diverse si ottengono distribuzioni finali diverse. La non-commutatività è una proprietà della geometria riemanniana curva di P: le proiezioni m-piatte in direzioni diverse non commutano perché la varietà P non è piatta.

La non-commutatività dell'attenzione ha conseguenze dirette per la cognizione sequenziale: l'ordine in cui gli stimoli ricevono attenzione determina la distribuzione finale delle credenze, non solo l'insieme degli stimoli attenzionati. Due sequenze dello stesso insieme di osservazioni, elaborate con la stessa maschera di attenzione ma in ordine diverso, producono distribuzioni di credenze diverse. Questo è il fondamento formale del primacy effect e del recency effect nella psicologia cognitiva: gli stimoli elaborati per primi e per ultimi modificano le credenze in modo asimmetrico perché la geometria di P è curva e le proiezioni non commutano.

II.5 La curvatura dello spazio delle credenze e la struttura dei bias

La curvatura della varietà P — misurata dal tensore di curvatura riemanniana associato alla metrica di Fisher — determina la struttura dei bias cognitivi che emergono dall'aggiornamento attenzionato parziale.

Il tensore di curvatura R della connessione e-piatta nello spazio esponenziale è identicamente zero: lo spazio delle famiglie esponenziali è e-piatto. Il tensore di curvatura della connessione m-piatta è in generale non nullo, e la sua struttura determina come le distribuzioni di credenze si "curvano" attorno ai punti di alta concentrazione.

Le regioni di P con alta curvatura m-piatta corrispondono a distribuzioni vicine al bordo del simplesso delle probabilità — distribuzioni fortemente concentrate su poche ipotesi. In quelle regioni, piccole variazioni nella maschera di attenzione producono grandi variazioni nella traiettoria dell'aggiornamento: il sistema è sensibile

all'attenzione. Le regioni con bassa curvatura corrispondono a distribuzioni diffuse: in quelle regioni, il sistema è robusto all'attenzione — variazioni nella maschera producono piccole variazioni nella traiettoria.

Questo produce la seguente struttura dei bias: un agente in stato di alta confidenza (distribuzione concentrata, alta curvatura) è altamente sensibile alla direzione dell'attenzione — piccole distorsioni nella maschera producono grandi deriva nella traiettoria. Un agente in stato di alta incertezza (distribuzione diffusa, bassa curvatura) è robusto alle distorsioni dell'attenzione. Il *confirmation bias* — la tendenza a cercare selettivamente le informazioni che confermano le credenze esistenti — è la struttura attenzionale che massimizza la sensibilità della propria traiettoria alle proprie credenze attuali: si concentra l'attenzione sulla regione di ipotesi già fortemente credute, sfruttando l'alta curvatura in quella regione per produrre aggiornamenti che rinforzano ulteriormente la concentrazione.

Formalmente, il confirmation bias è la maschera $A_t^{confirm}(h) \propto \pi_t(h)$: la maschera proporzionale alla distribuzione corrente. Sotto questa maschera, il gradiente attenzionato è:

$$\nabla_P^{A^{confirm}} \mathcal{F}(\pi_t ; o)(h) \propto \pi_t(h) \cdot [\log P(o|h) - \mathbb{E}_{\pi_t}[\log P(o|\cdot)]]$$

il gradiente bayesiano pesato dalla probabilità a priori. Questo aggiornamento amplifica le variazioni nelle ipotesi già credute e sopprime le variazioni nelle ipotesi già scartate, producendo un restringimento progressivo della distribuzione anche in assenza di evidenze genuine.

II.6 Il gradiente precondizionato e la struttura dell'expertise attenzionale

Il gradiente naturale attenzionato del paragrafo II.3 richiede il calcolo di $G(\pi_t)^{-1}$ — l'inversa della matrice di Fisher, che può essere computazionalmente costosa. Algoritmi pratici usano approssimazioni: il gradiente precondizionato con una matrice di precondizionamento $M_t \approx G(\pi_t)$:

$$\tilde{\nabla}_P^{A,M} \mathcal{F}(\pi_t ; o) = M_t^{-1} \cdot \nabla_P^A \mathcal{F}(\pi_t ; o)$$

La scelta della matrice di precondizionamento M_t è la formalizzazione del processo di costruzione dell'expertise: un esperto nel dominio D ha una matrice di precondizionamento M_t^{expert} che approssima $G(\pi_t)$ molto più fedelmente di un novizio. L'esperto "sa dove guardare" — ha un modello della curvatura dello spazio delle credenze nel proprio dominio che amplifica i segnali rilevanti e sopprime il rumore.

La connessione con l'expertise percettiva del Capitolo VI della seconda monografia è la seguente. Il medico che nomina le diagnosi sviluppa, attraverso il termine f_{lang} del sistema dinamico accoppiato, una trasformazione metrica dello spazio percettivo

$\mathbb{R}^n$_{visivo} che amplifica le distinzioni categorialmente rilevanti. Nel framework del gradiente naturale, questo corrisponde alla costruzione di una matrice di precondizionamento M_t^{medico} che è l'inversa della matrice di Fisher dello spazio delle ipotesi diagnostiche: le ipotesi diagnostiche rilevanti hanno gradiente amplificato, quelle irrilevanti hanno gradiente soppresso. L'expertise è la costruzione di un buon precondizionatore per il problema di aggiornamento bayesiano nel proprio dominio.

Questa connessione unifica la teoria dell'attenzione con la teoria dell'expertise della seconda monografia: l'esperto non è semplicemente chi ha più conoscenza, ma chi ha un sistema attenzionale (un precondizionatore) che accelera l'aggiornamento nelle dimensioni rilevanti e rallenta quello nelle dimensioni irrilevanti, sfruttando la geometria riemanniana dello spazio delle credenze nel proprio dominio.

Capitolo III — Il controllo dell'attenzione come controllo dell'evoluzione delle credenze

III.1 La struttura formale dell'influenza epistemica

Il Capitolo I ha stabilito che l'attenzione selettiva produce aggiornamenti delle credenze che divergono sistematicamente dall'ottimo bayesiano. Il Capitolo II ha precisato la geometria di questa divergenza nello spazio riemanniano delle distribuzioni. Il Capitolo III affronta la conseguenza politica di questa struttura: chiunque abbia la capacità di modificare la maschera di attenzione A_t di un agente — di determinare su quale porzione del gradiente bayesiano l'agente applica la propria elaborazione — ha la capacità di controllare la traiettoria delle credenze di quell'agente.

Si definisce formalmente l'*influenza epistemica*. Un agente esterno E ha influenza epistemica sull'agente bersaglio T se può produrre segnali $s \in \mathcal{S}$ che modificano la maschera di attenzione A_t^T di T. L'influenza epistemica è il controllo della selezione del gradiente del bersaglio. Il *potere epistemico* di E su T è:

$\Pi(E, T) = \sup_{s \in \mathcal{S}} D_{KL}(\pi_T^{s} \| \pi_T^{\varnothing})$

la massima divergenza KL che E può produrre tra le credenze di T con il segnale s e le credenze che T avrebbe in assenza di intervento — misurata attraverso la norma del displacement massimale nello spazio delle credenze. Un agente con alto potere epistemico può produrre grandi spostamenti nelle credenze del bersaglio; uno con basso potere epistemico può produrre solo piccoli spostamenti.

III.2 Tre operatori sull'attenzione: distruzione, cattura, sostituzione

Si identificano tre meccanismi di influenza epistemica con strutture formali distinte.

Il primo è la *distruzione dell'attenzione*: il produttore del segnale genera un flusso di stimoli ad alta salienza ma basso contenuto informazionale genuino, saturando la capacità attenzionale C del bersaglio con input che non portano il gradiente bayesiano rilevante. Formalmente, l'agente esterno aumenta il denominatore del problema di ottimizzazione dell'attenzione: la capacità C rimane la stessa, ma il numero di stimoli che competono per essa aumenta, riducendo la quota attenzionale disponibile per ogni singolo stimolo rilevante. Il bersaglio subisce una riduzione del η_eff per gli stimoli informativamente rilevanti non perché sia meno capace ma perché la sua capacità è stata monopolizzata da stimoli irrilevanti ad alta salienza.

La distruzione dell'attenzione è l'operatore formale dell'*information overload* nel mercato dell'informazione moderno: la sovrabbondanza di stimoli non aumenta la capacità di aggiornamento delle credenze ma la riduce, perché ogni stimolo riceve una quota attenzionale inferiore a quella ottimale. La convergenza verso credenze accurate si rallenta come $\propto C/(C + C_rumore)$ dove C_rumore è la capacità assorbita dal rumore: aumentare il rumore degrada l'aggiornamento anche senza modificarne la direzione.

Il secondo meccanismo è la *cattura dell'attenzione*: l'agente esterno produce segnali che ridirigono sistematicamente la maschera A_t^T verso un sottoinsieme di ipotesi $H_A \subset H$ predeterminato — tipicamente il sottoinsieme delle ipotesi favorevoli alla narrativa del produttore. Formalmente, si sostituisce la maschera ottimale $A_t^{T,}$ *con la maschera manipolata* $A_t^{T,manip}(h) = a \cdot A_t^{T,}(h) + (1-\alpha) \cdot A^{E}(h)$ dove A^{E} è la maschera favorevole al produttore e $\alpha \in [0,1]$ misura il grado di resistenza del bersaglio alla manipolazione. Per $\alpha \to 0$ il bersaglio elabora esclusivamente il sottoinsieme ipotetico voluto dal produttore.

Il terzo meccanismo è la *sostituzione del gradiente*: l'agente esterno fornisce al bersaglio non il segnale grezzo $o_\{t+1\}$ ma una trasformazione del segnale $\Psi(o_\{t+1\})$ che distorce il gradiente bayesiano prima che il bersaglio lo elabori. La trasformazione Ψ può amplificare certe componenti del gradiente (enfatizzare certi aspetti dell'osservazione), sopprimerne altre (mascherare certi aspetti), o inserire componenti spurie (produrre l'impressione di un gradiente verso ipotesi false). La sostituzione del gradiente è il meccanismo formale della *disinformazione*: il bersaglio applica correttamente la propria funzione di aggiornamento, ma il segnale che riceve è stato manipolato prima di raggiungerlo.

III.3 La propaganda come campo di distorsione del gradiente

La propaganda — nel senso tecnico più ampio, non necessariamente peggiorativo — è l'insieme dei meccanismi attraverso cui produttori di segnali modificano la distribuzione di attenzione di una collettività di agenti in modo pianificato e sistematico. La formalizzazione della propaganda come campo di distorsione del gradiente permette di classificare i suoi meccanismi con precisione.

Sia $\{\pi_i^t\}\{i \in N\}$ *la distribuzione delle credenze degli* N *agenti di una collettività al tempo* t. *Un programma propagandistico è una sequenza di segnali* $\{s_\tau\}\{\tau \leq t\}$ progettata per portare

la distribuzione aggregata $\varrho_t = (1/N)\ \Sigma_i\ \pi_i^t$ verso un target Π^{target} — la distribuzione di credenze desiderata dal produttore. Il programma ottimale risolve:

$$\{s_\tau^*\}_{\tau \leq T} = \operatorname{argmin}_{\{s_\tau\}} D_{KL}(\varrho_T \,\|\, \Pi^{target})$$

soggetto ai vincoli sulla struttura dei segnali producibili (budget, canali disponibili, credibilità del mittente). Questo è un problema di controllo ottimale nello spazio delle distribuzioni di distribuzioni — un problema di secondo ordine nella struttura della torre bayesiana gerarchica della terza monografia.

Il gradiente del funzionale obiettivo rispetto al segnale s_τ è:

$$\delta/\delta s\ D_{KL}(\varrho_T \,\|\, \Pi^{target}) = (1/N)\ \Sigma_i\ \partial D_{KL}(\pi_i^T \,\|\, \Pi^{target})/\partial s_\tau$$

la media sui bersagli della sensibilità delle loro credenze finali al segnale al tempo τ. Il segnale ottimale al tempo τ è quello che riduce massimalmente la divergenza media dal target per unità di costo del segnale. Questo dipende dalla distribuzione attuale delle credenze ϱ_τ: il segnale ottimale è diverso nelle diverse fasi della campagna, e la struttura ottimale della sequenza temporale dei segnali è una politica di controllo adattativa.

III.4 La persuasione legittima e la manipolazione: una distinzione formale

Il framework del gradiente bayesiano permette di proporre una distinzione formale — non soltanto normativa — tra persuasione epistemicamente legittima e manipolazione. La distinzione non è semplice da tracciare, e molte teorie etiche della comunicazione la tracciano in modo intuitivo (rispetto dell'autonomia, consenso informato, verità del contenuto). Il framework formale offre un criterio strutturale più preciso.

Si dice che un'influenza epistemica è *legittima* se il segnale prodotto dall'agente esterno corrisponde a un'osservazione genuina — un campione dal processo generativo reale che governa le ipotesi H. Formalmente, il segnale s è legittimo se esiste un processo generativo P^{real} tale che $s \sim P^{real}(\cdot \mid h)$ per qualche $h \in H$. In questo caso, il gradiente prodotto nel bersaglio punta nella direzione di h — è un gradiente verso la verità, anche se l'agente esterno ha scelto di rivelare questa particolare osservazione anziché un'altra.

Si dice che un'influenza epistemica è *manipolatoria* se il gradiente che induce nel bersaglio è sistematicamente scollegato dal processo generativo reale. Formalmente, la manipolazione produce un gradiente:

$$\nabla_{P^{manip}}\, \mathcal{F}(\pi_t ; s)(h) \neq \mathbb{E}_{P^{real}}[\nabla_P\, \mathcal{F}(\pi_t ; o)(h)]$$

un gradiente che punta in una direzione sistematicamente distante dalla direzione del gradiente bayesiano ottimale sotto il processo generativo reale. La manipolazione è la produzione di un gradiente spurio: un vettore nello spazio delle credenze che non corrisponde ad alcun segnale genuino.

Questa distinzione formale è più sottile della semplice opposizione "vero/falso". Un'affermazione vera selettivamente scelta per le sue conseguenze attenzionali può essere manipolatoria nel senso formale: se la selezione produce un gradiente che punta sistematicamente lontano dalla distribuzione ottimale, l'influenza è manipolatoria anche se ogni singola affermazione è vera. La *cherry-picking* è la struttura formale della manipolazione attraverso vere affermazioni: si selezionano osservazioni genuine i cui gradienti, sommati, puntano verso Π^{target} anche quando il gradiente integrale del processo generativo reale punta altrove.

III.5 L'economia dell'attenzione come campo di forze sul gradiente collettivo

I mercati dell'informazione — piattaforme digitali, media, sistemi educativi — sono meccanismi che determinano quale distribuzione di maschere di attenzione si stabilisce nella collettività. Il produttore che riesce a fare sì che la propria narrativa riceva attenzione da una frazione Φ della collettività esercita un potere epistemico proporzionale a $\Phi \cdot |\nabla_P^{manip}\mathcal{F}|$: la forza del proprio gradiente per la quota di agenti che lo elaborano.

Si formalizza il mercato dell'attenzione come un campo di forze sul gradiente collettivo. Ogni produttore di segnali P_k genera un campo di distorsione $\Delta_k: H \to \mathbb{R}$ che rappresenta il displacement che i suoi segnali producono nel gradiente bayesiano degli agenti che li ricevono. Il campo di forze totale è la sovrapposizione pesata dei displacement individuali:

$$\Delta^{tot}(h) = \Sigma_k \, w_k \cdot \Delta_k(h)$$

dove $w_k \in [0,1]$ è la quota attenzionale che il produttore k riesce a catturare dalla collettività, con $\Sigma_k \, w_k \leq 1$. Il campo Δ^{tot} è il gradiente aggiuntivo che il sistema informativo imprime sulle credenze collettive oltre al gradiente generato dall'esperienza diretta. L'equilibrio del mercato dell'informazione è la distribuzione $\{w_k^*\}$ che massimizza l'utilità aggregata dei produttori soggetta ai vincoli della competizione per l'attenzione.

Il risultato fondamentale è che l'equilibrio del mercato dell'attenzione non coincide in generale con l'equilibrio epistemicamente ottimale: la distribuzione $\{w_k^*\}$ che massimizza il profitto dei produttori non è quella che massimizza la calibrazione delle credenze degli agenti bersaglio. La ragione è che i segnali ad alta salienza — quelli che catturano attenzione — non sono necessariamente quelli con il più alto contenuto informazionale genuino. La competizione nel mercato dell'attenzione seleziona i

produttori che ottimizzano per la cattura dell'attenzione (massimizzazione di w_k) anziché per l'accuratezza del gradiente (massimizzazione della calibrazione).

Capitolo IV — Economia politica formale dell'attenzione collettiva

IV.1 I mercati dell'attenzione come meccanismi di aggregazione

Nel Capitolo V della terza monografia, la metacognizione collettiva è stata trattata come il problema dell'aggregazione dei meta-modelli individuali $\{\pi^0_i\}$, soggetta ai teoremi di impossibilità di Arrow e Hylland-Zeckhauser. Il presente capitolo estende quella analisi al livello dell'attenzione: il problema di aggregare le maschere di attenzione individuali {A_t^i} in un campo attenzionale collettivo A_t^C — il meccanismo che determina quale porzione del gradiente bayesiano collettivo viene effettivamente elaborata.

Un *meccanismo di aggregazione dell'attenzione* è una funzione Agg : {A_t^i}_{i ∈ N} → A_t^C che a ogni profilo di maschere individuali associa una maschera collettiva. Analogamente all'aggregazione delle preferenze di Arrow, si possono enunciare proprietà desiderabili per Agg:

(U) *Dominio illimitato*: Agg è definita per tutti i profili di maschere individuali.

(P) *Principio di Pareto attenzionale*: se tutti gli agenti assegnano A_t^i(h) = 1 a un'ipotesi h, allora A_t^C(h) = 1.

(IIA-A) *Indipendenza delle ipotesi irrilevanti*: A_t^C(h) dipende solo dai valori {A_t^i(h)}_i e non dai valori su altre ipotesi.

(ND-A) *Non-dittatura attenzionale*: nessun agente i determina A_t^C = A_t^i indipendentemente dalle maschere degli altri.

(C) *Continuità*: piccole variazioni nelle maschere individuali producono piccole variazioni nella maschera collettiva.

Teorema di impossibilità per l'aggregazione delle maschere di attenzione: Per spazi di ipotesi H continui e insiemi di agenti N con |N| ≥ 2, nessuna funzione di aggregazione Agg soddisfa simultaneamente (U), (P), (IIA-A), (ND-A), e (C).

La dimostrazione è una riduzione al teorema di Arrow. Si discretizza H in un insieme finito $\{h_1, ..., h_m\}$ di ipotesi: il problema di aggregazione delle maschere diventa un problema di aggregazione di funzioni booleane sull'insieme delle ipotesi, che è equivalente a un problema di aggregazione delle preferenze con preferenze su $\{h_1, ..., h_m\}$. Il teorema di Arrow si applica direttamente producendo l'impossibilità.

Il teorema afferma: non esiste un meccanismo di aggregazione dell'attenzione collettiva che rispetti l'autonomia attenzionale di tutti gli agenti, sia non-dittatoriale, e produca maschere collettive continue. I mercati dell'attenzione sono sempre, in senso formale, dittatoriali in qualche dimensione dello spazio ipotetico: esiste sempre una regione di H la cui attenzione collettiva è determinata da un sottoinsieme ristretto di agenti anziché dalla media ponderata di tutti.

IV.2 La struttura del potere epistemico: concentrazione e distribuzione

Il teorema di impossibilità dell'Appendice A della terza monografia — il risultato di Arrow come istanza di Lawvere — si applica qui con una nuance importante. Nel contesto attenzionale, la "dittatura" non è necessariamente esercitata da un singolo agente: può essere distribuita su un insieme di agenti che condividono la stessa maschera attenzionale, producendo una *oligarchia attenzionale*.

Si definisce la *concentrazione del potere epistemico* come:

$$\Gamma(A^C) = \min_{S \subseteq N : A^C = \mathrm{Agg}(\{A^i\}_{i \in S})} |S| / N$$

la più piccola frazione della collettività che determina completamente la maschera collettiva. Un meccanismo con $\Gamma = 1/N$ è completamente democratico (ogni agente ha peso uguale); un meccanismo con $\Gamma \to 0$ è oligarchico (una piccola frazione determina l'attenzione collettiva).

I mercati dell'attenzione basati sulla competizione per la salienza tendono a produrre Γ piccolo: il meccanismo di mercato seleziona i produttori più efficaci nel catturare l'attenzione, che tipicamente sono quelli con risorse (budget pubblicitario, infrastruttura tecnica, posizionamento algoritmico) molto superiori alla media. La concentrazione del potere epistemico in questi mercati è l'effetto della dinamica competitiva che porta a equilibri con $\Gamma << 1$.

La misura di concentrazione Γ è correlata con una misura di diversità epistemica della collettività:

$$\mathscr{D}(\rho_T) = \Sigma_i H(\pi_i^T) / (N \cdot H(\Pi^{\mathrm{uniform}}))$$

la diversità media delle credenze individuali normalizzata all'entropia massima. Per $\Gamma \to 0$ (alta concentrazione), la diversità tende a zero: tutti gli agenti convergono alle credenze determinate dalla maschera dell'oligarchia attenzionale. Per $\Gamma \to 1$ (bassa concentrazione), la diversità può essere alta: agenti diversi mantengono maschere diverse e converging a credenze diverse.

Il trade-off tra concentrazione epistemica ed efficienza dell'aggiornamento è il nucleo dell'economia politica formale dell'attenzione. La concentrazione riduce i costi di coordinamento (un'unica maschera collettiva è computazionalmente più efficiente di

N maschere diverse) ma riduce la diversità epistemica (la collettività è meno capace di esplorare lo spazio delle ipotesi in modo ampio). La diversità epistemica è il valore della "perplessità collettiva" — la capacità della collettività di mantenere ipotesi diverse aperte e di aggiornare in direzioni diverse in risposta a segnali diversi. Una collettività con alta diversità epistemica è più robusta alla manipolazione: è più difficile costruire un campo di distorsione del gradiente che sposta in modo uniforme tutte le credenze.

IV.3 Il teorema di Sen per l'attenzione: libertà epistemica e aggregazione

Il teorema di Sen sull'impossibilità del liberale paretiano trova un'istanza diretta nel contesto attenzionale. Il *liberalismo epistemico* è la condizione che ogni agente abbia il diritto di determinare la propria maschera di attenzione su almeno una coppia di ipotesi — la libertà di decidere a cosa prestare attenzione nel proprio dominio di competenza.

Si enuncia formalmente il *teorema di impossibilità del liberale epistemico*: per qualsiasi meccanismo di aggregazione Agg che soddisfi il principio di Pareto attenzionale (P) e il liberalismo epistemico minimo, esistono profili di maschere individuali per cui Agg produce una maschera collettiva incoerente — che assegna attenzione insufficiente a ipotesi che tutti gli agenti vorrebbero attenzionare o eccessiva a ipotesi che nessun agente vuole attenzionare.

La dimostrazione segue lo schema di Sen: si costruisce un esempio con tre agenti e tre ipotesi in cui le condizioni di Pareto e di liberalismo minimo sono mutuamente contraddittorie per qualsiasi meccanismo di aggregazione. Il caso esplicito: h_1 = "l'ipotesi che X sia vero", h_2 = "l'ipotesi che Y sia conseguenza di X", h_3 = "l'ipotesi che Y sia direttamente osservabile". L'agente 1 vuole attenzionare h_1 prima di h_2; l'agente 2 vuole attenzionare h_2 prima di h_3; tutti e tre vogliono attenzionare h_3 più di h_1. Il principio di Pareto richiede $A^C(h_3) > A^C(h_1)$; il liberalismo di agente 1 richiede $A^C(h_1) > A^C(h_2)$; il liberalismo di agente 2 richiede $A^C(h_2) > A^C(h_3)$. La composizione produce $A^C(h_3) > A^C(h_1) > A^C(h_2) > A^C(h_3)$: una contraddizione.

Il teorema di Sen per l'attenzione afferma che non esiste un modo di aggregare le preferenze attenzionali individuali che rispetti simultaneamente l'autonomia delle scelte attenzionali personali e l'efficienza collettiva dell'aggiornamento. Qualsiasi sistema informativo deve scegliere tra rispettare la libertà attenzionale individuale e ottimizzare la propagazione dei segnali rilevanti: non può fare entrambe le cose simultaneamente per tutti i profili di preferenze attenzionali.

IV.4 Meccanismi di secondo grado: l'attenzione sull'attenzione

I meccanismi di influenza epistemica di primo grado — quelli che modificano direttamente la maschera di attenzione A_t^T del bersaglio — sono soggetti alla resistenza del bersaglio: un agente consapevole della manipolazione può adottare contromisure, ignorare certi segnali, aggiornare le proprie aspettative sulla fonte.

I meccanismi di secondo grado sono più sottili e più difficili da resistere: anziché modificare la maschera di attenzione direttamente, modificano la *struttura di selezione della maschera* — la funzione che determina quale maschera A_t il bersaglio adotta in risposta a determinati stimoli. Formalmente, il meccanismo di secondo grado agisce sul livello π^1 della torre metacognitiva: anziché produrre input che modificano π^0 attraverso la maschera corrente, modifica la credenza di secondo ordine del bersaglio sulla propria maschera ottimale.

Un meccanismo di secondo grado efficace porta il bersaglio a credere che la propria maschera distorta sia ottimale — che la direzione di attenzione indotta dal produttore sia quella che il bersaglio stesso sceglierebbe razionalmente. Questo è il meccanismo della *falsa coscienza* nella sua formulazione formale: il bersaglio adotta la maschera del produttore non perché non abbia scelta ma perché il suo meta-modello della propria maschera ottimale è stato modificato dal produttore attraverso segnali di secondo ordine.

L'influenza epistemica di secondo grado è formalmente equivalente a un'azione sul sistema dinamico accoppiato della seconda monografia al livello del termine g_comm — il termine che governa l'aggiornamento della competenza linguistica in risposta all'input comunicativo. Modificare g_comm attraverso segnali che ridefiniscono il vocabolario disponibile per ragionare sull'attenzione — per ragionare su cosa merita considerazione, su quali argomenti sono "seri" e quali non lo sono — modifica il punto fisso del sistema dinamico accoppiato verso una configurazione in cui certi tipi di attenzione sono strutturalmente incoraggiati e altri sono strutturalmente soppressi.

IV.5 La polarizzazione come equilibrio di Nash del campo attenzionale

La polarizzazione epistemica collettiva — la tendenza di una società a convergere verso credenze bimodali con due cluster distanti anziché una distribuzione unimodale — è un fenomeno empirico ben documentato che il framework del gradiente attenzionale illumina formalmente.

Si considera un sistema collettivo di N agenti che aggiornano le proprie credenze attraverso una combinazione di osservazione diretta (gradiente bayesiano da esperienze personali) e osservazione mediata (gradiente bayesiano filtrato attraverso il sistema informativo). Il campo attenzionale del sistema informativo — la distribuzione delle maschere che i diversi produttori di informazione impongono ai diversi segmenti della collettività — determina quali segnali ricevono attenzione in quali segmenti.

Un equilibrio di Nash del campo attenzionale è un profilo di maschere {A_t^{k,*}} tale che nessun produttore k può aumentare il proprio potere epistemico (la propria quota di attenzione w_k) modificando unilateralmente la propria maschera. La struttura degli equilibri di questo gioco dipende dalla funzione di payoff dei produttori: se i produttori massimizzano l'attenzione catturata (click, visualizzazioni)

piuttosto che l'accuratezza del gradiente prodotto, l'equilibrio è la selezione per la salienza.

Il risultato cruciale è che gli equilibri di Nash del mercato dell'attenzione con payoff basati sulla salienza sono tipicamente polarizzanti: la competizione per l'attenzione seleziona contenuti che amplificano le differenze tra segmenti della collettività, perché i segnali che producono forte risposta attenzionale sono spesso quelli che confermano intensamente le credenze di un gruppo (salienza attraverso la conferma) o che producono forte risposta emotiva negativa verso le credenze dell'altro gruppo (salienza attraverso la minaccia).

Formalmente: l'equilibrio di Nash con payoff-salienza produce un campo attenzionale polarizzante se la funzione di salienza $\sigma(s, \pi_i)$ — la probabilità che l'agente i con credenze π_i presti attenzione al segnale s — è maggiore per segnali s tali che $\nabla_P{}^s \mathcal{F}(\pi_i)$ è allineato con π_i (conferma) o ortogonale a π_i nelle dimensioni di alta credenza (minaccia) rispetto ai segnali con gradiente verso la distribuzione bayesiana ottimale π_i^{Bayes}. Questa condizione è soddisfatta da qualsiasi funzione di salienza che privilegia la coerenza emotiva o la valenza affettiva rispetto al contenuto informazionale puro.

Capitolo V — L'auto-attenzione e il punto fisso diagonale: chiusura del ciclo

V.1 Il problema dell'auto-attenzione

I capitoli precedenti hanno trattato l'attenzione come un meccanismo che processa input esterni — osservazioni dal mondo, segnali da altri agenti. Ma il sistema attenzionale può rivolgere se stesso su se stesso: il meccanismo che seleziona i gradienti da applicare è esso stesso un processo cognitivo soggetto all'attenzione. Quando l'attenzione si rivolge sulla propria struttura — quando il sistema elabora segnali sulla propria maschera attenzionale, sulla propria storia di aggiornamenti, sulla propria struttura di credenze — si genera la struttura dell'*auto-attenzione*.

L'auto-attenzione non è una curiosità marginale: è il meccanismo centrale della coscienza riflessiva. Il momento in cui un sistema cognitivo diventa consapevole della propria struttura attenzionale è il momento della auto-riflessione metacognitiva — il passaggio dal livello π^0 al livello π^1 della torre bayesiana della terza monografia. La connessione tra attenzione e metacognizione è precisa: la metacognizione è l'auto-attenzione che ha prodotto credenze stabili sulla propria struttura attenzionale.

Formalmente, si definisce l'*operatore di auto-attenzione* $\mathcal{S}$ come il caso speciale dell'operatore di attenzione $\mathcal{A}_t$ in cui lo spazio delle ipotesi H è lo spazio delle maschere attenzionali possibili: $H_A = \{A : H \to [0,1]\}$. Le ipotesi sono ora maschere — il sistema forma credenze su quale sia la propria maschera corrente — e

l'osservazione o_{t+1} è un'osservazione del comportamento del proprio sistema attenzionale (cosa ha elaborato, cosa ha trascurato, quali errori ha commesso).

V.2 Il punto fisso dell'auto-attenzione: la struttura dell'identità cognitiva

Un punto fisso dell'operatore di auto-attenzione è una maschera A* tale che il sistema con credenza π* sull'insieme delle maschere, aggiornato attraverso l'osservazione del proprio comportamento attenzionale, produce un aggiornamento che conserva A* come la maschera più creduta:

$\mathcal{I}[\pi^*] = \pi^*$

dove $\pi^* \in P(H_A)$ è concentrata intorno a A*. Il punto fisso dell'auto-attenzione è la struttura dell'*identità cognitiva*: il sistema ha una maschera stabile che si auto-riproduce attraverso l'osservazione del proprio comportamento — osserva che usa A*, aggiorna la credenza che la propria maschera sia A*, il che porta a continuare a usare A*.

Il punto fisso dell'auto-attenzione ha due tipi di stabilità. Il punto fisso *stabile* è quello verso cui il sistema converge da condizioni iniziali vicine: una perturbazione della maschera produce un aggiornamento che riporta verso A*. Questo è la struttura formale della robustezza dell'identità: il sistema "resistente" alle perturbazioni attenzionali esterne, che tendono a riportare la maschera verso il punto fisso. Il punto fisso *instabile* è quello da cui il sistema si allontana sotto piccole perturbazioni: perturbazioni della maschera si amplificano anziché dissiparsi. Questo è la struttura formale della *crisi dell'identità*: un sistema in un punto fisso instabile è suscettibile a transizioni verso punti fissi alternativi sotto perturbazioni sufficientemente grandi.

V.3 Il lemma diagonale per l'auto-attenzione

Il lemma diagonale di Gödel-Carnap del Capitolo VIII della terza monografia si applica direttamente all'auto-attenzione, producendo un risultato di impossibilità della struttura familiare.

Si definisce la *mappa di auto-attenzione* come la funzione:

$\Psi : H_A \to P(H_A)$
$\Psi(A)(A') = P(\text{sistema con maschera A usi A' come passo successivo})$

che a ogni maschera associa la distribuzione sulla propria maschera successiva — la distribuzione sul comportamento del sistema attenzionale visto da se stesso. Il lemma diagonale per l'auto-attenzione afferma: se Ψ è sufficientemente espressiva — se può rappresentare qualsiasi relazione tra maschere correnti e future — allora per ogni operatore $T : P(H_A) \to P(H_A)$ esiste una maschera A* tale che:

$\Psi(A^*) = T(\Psi(A^*))$

A* è la maschera che si auto-attribuisce esattamente la distribuzione prodotta dall'operatore T applicato alla propria distribuzione. Per T = identità, il punto fisso è la maschera che si auto-riproduce esattamente — il punto fisso stabile dell'identità. Per T = negazione (nella versione discreta), il punto fisso è impossibile — analogo alla formula di Gödel — producendo una contraddizione nell'auto-attribuzione.

La conseguenza metacognitiva è la seguente. Un sistema attenzionale sufficientemente espressivo — uno che può formarsi credenze su qualsiasi relazione tra la propria maschera e il proprio comportamento — è soggetto al lemma diagonale: esiste sempre una struttura di auto-attribuzione che non ha soluzione coerente. Il sistema non può avere, simultaneamente, credenze complete e coerenti sulla propria struttura attenzionale. Questo è il limite dell'auto-conoscenza attenzionale: non è un limite di sofisticazione ma un limite strutturale che il lemma diagonale rende formalmente preciso.

V.4 La struttura dei bacini di attrazione e la topologia delle identità cognitive

I punti fissi stabili dell'auto-attenzione formano un insieme discreto nel paesaggio delle maschere attenzionali possibili. Ogni punto fisso stabile A_*k ha un bacino di attrazione B_k* $\subset$ *H_A — l'insieme delle maschere iniziali da cui il sistema converge a A*_k sotto la dinamica dell'auto-attenzione. La struttura dei bacini di attrazione è la *topologia delle identità cognitive*.

Le frontiere tra bacini di attrazione — le *separatrici* — sono le regioni dello spazio delle maschere da cui il sistema può convergere a identità diverse. Un sistema che si trova vicino a una separatrice è strutturalmente instabile: piccole perturbazioni attenzionali (brevi esposizioni a segnali che inducono una maschera diversa) possono spingerlo verso un bacino diverso. Le trasformazioni della terapia cognitiva, della conversione religiosa, della crisi dell'identità, e della riqualificazione professionale sono tutti casi di sistemi che attraversano una separatrice e convergono verso un bacino di attrazione diverso — un nuovo punto fisso dell'auto-attenzione con struttura A_*new strutturalmente diversa da A*_old.

La struttura topologica delle separatrici determina quali transizioni di identità sono possibili e con quale resistenza. Per sistemi con un singolo bacino di attrazione globale (punto fisso globalmente stabile), tutte le perturbazioni si dissipano: l'identità è robusta a qualsiasi intervento. Per sistemi con molti bacini di attrazione (paesaggi energetici rugosi), le perturbazioni possono produrre transizioni irreversibili verso attrattori locali: l'identità è fragile rispetto a certi tipi di intervento e robusta rispetto ad altri.

V.5 La misura invariante e la distribuzione stazionaria delle credenze

In un sistema cognitivo con dinamica stocastica — aggiornamento bayesiano in risposta a un processo di osservazioni stocastico O_t — il punto fisso deterministico dell'auto-attenzione si generalizza in una *misura invariante* μ* sullo spazio P(H_A):

$$\Phi_\tau^* \mu^* = \mu^*$$

dove Φ_τ è il semigruppe dell'operatore di auto-attenzione. La misura invariante è la distribuzione stazionaria delle credenze attenzionali: la distribuzione verso cui il sistema converge a lungo termine indipendentemente dalle condizioni iniziali, se la dinamica è ergodica.

Per sistemi con ergodicicità (ogni regione dello spazio delle credenze è visitata con probabilità positiva sotto la dinamica stocastica), la misura invariante è unica e ha un'interpretazione precisa: è la distribuzione delle credenze che il sistema mantiene "in media" nel lungo termine, pesata dal tempo che passa in ogni regione dello spazio. La misura invariante dell'auto-attenzione è la struttura statistica dell'identità cognitiva di lungo termine.

Il risultato che connette i Capitoli III e V è il seguente. Un produttore di segnali che vuole modificare stabilmente l'identità cognitiva di un bersaglio deve modificare la misura invariante μ* — deve spostare il punto fisso dell'auto-attenzione anziché produrre semplici spostamenti temporanei della maschera. Questo richiede segnali che modificano il termine di auto-attribuzione Ψ(A) — la credenza del sistema sulla propria struttura — anziché soltanto la maschera corrente A_t. Il controllo dell'identità è più profondo e più resistente del controllo dell'attenzione momentanea: richiede un'influenza epistemica di secondo grado che modifichi il punto fisso dell'auto-attenzione.

V.6 Chiusura del ciclo: la struttura diagonale nella quarta monografia

Il filo diagonale che attraversa le quattro monografie trova qui la propria chiusura. La struttura di punto fisso impossibile prodotta dal lemma diagonale — che nelle prime tre monografie si manifesta come malinteso irriducibile, come incomprimibilità del substrato procedurale, come incompletezza dell'auto-conoscenza formale — si manifesta nella quarta come impossibilità dell'auto-attenzione completa.

Il sistema che tenta di applicare attenzione completa alla propria struttura attenzionale produce il punto fisso impossibile: la maschera che si auto-attribuisce esattamente il comportamento della propria negazione. Questo non è il problema della coscienza — non spiega perché ci sia qualcosa che si provi ad essere un sistema attenzionale — ma lo localizza con precisione: la struttura formale che la coscienza riflessiva deve navigare è la struttura del lemma diagonale applicato all'auto-attenzione, con le sue regioni di stabilità (i punti fissi) e le sue frontiere di impossibilità (le separatrici).

Le quattro monografie convergono su un risultato unificato: la cognizione come sistema che si auto-rappresenta è soggetta a limiti strutturali irriducibili — nella

comunicazione (malinteso), nella compressione (substrato), nell'auto-conoscenza (metacognizione), nell'auto-attenzione — la cui forma comune è lo schema diagonale di Lawvere. Non c'è un livello della cognizione a cui il sistema si libera da questi limiti. Ma i limiti non producono impotenza: producono la struttura specifica entro cui la cognizione opera, con i suoi punti fissi, i suoi attrattori, i suoi confini di impossibilità che sono al tempo stesso i contorni dell'identità.

Capitolo VI — Precision weighting, active inference e i residui del framework di Friston

VI.1 Perché la connessione è necessaria senza ripetere il saggio

Il principio di energia libera di Friston e il framework dell'active inference sono trattati altrove nel ciclo in modo autonomo. Quello che manca in questa monografia non è la derivazione del principio — che il lettore trova nel saggio dedicato — ma la traduzione precisa di due oggetti specifici di quel framework nel vocabolario del gradiente attenzionato: la matrice di precisione Π come caso particolare del precondizionatore attenzionale M_t, e la gerarchia delle precisioni come struttura che connette l'active inference con la torre bayesiana gerarchica della terza monografia. Questa sezione sviluppa esclusivamente quei due ponti, senza duplicare la derivazione del funzionale di energia libera.

VI.2 La precision weighting come precondizionamento attenzionale

Nel framework variazionale di Friston, le credenze dell'agente su uno stato nascosto h sono distribuzioni di probabilità approssimate $q(h) \in P$ parametrizzate — tipicamente gaussiane con media μ e varianza Σ^{-1} dove Σ è la matrice di precisione. L'aggiornamento che minimizza la variational free energy rispetto ai parametri della distribuzione approssimata produce:

$$d\mu/dt = -\partial F/\partial \mu = \Sigma \cdot (\mu_prior - \mu) + J^T \cdot \Sigma_o \cdot \varepsilon_o$$

dove $\varepsilon_o = o - g(\mu)$ è l'errore di predizione dell'osservazione, g è la funzione generativa, $J = \partial g/\partial \mu$ è il Jacobiano, e Σ_o è la matrice di precisione sull'osservazione. L'equazione dice: l'aggiornamento della media è proporzionale alla somma di un termine di ritorno al prior (ponderato dalla precisione del prior Σ) e di un termine di correzione dall'osservazione (ponderato dalla precisione dell'osservazione Σ_o).

La connessione con il gradiente naturale attenzionato del Capitolo II è immediata. Si scriva l'aggiornamento nella forma:

$$d\mu/dt = -G(q)^{-1} \cdot \nabla_\mu F \quad \text{con} \quad G(q)^{-1} = (\Sigma + J^T \Sigma_o J)^{-1}$$

La matrice $(\Sigma + J^T \Sigma_o J)$ è esattamente la matrice di Fisher della distribuzione congiunta prior-verosimiglianza valutata in μ: è il precondizionatore naturale per l'aggiornamento bayesiano variazionale. L'attenzione, in active inference, è la modulazione della matrice di precisione Σ_o: aumentare $\Sigma_o(h, h)$ equivale ad aumentare il peso dell'errore di predizione nella direzione dell'ipotesi h, il che è esattamente il caso $A_t(h) = \Sigma_o(h,h)/\max_k \Sigma_o(k,k)$ — la normalizzazione della precisione sull'osservazione è la maschera di attenzione.

Formalmente, si stabilisce la corrispondenza:

$A_t(h) \longleftrightarrow \Sigma_o(h, h) / Z_\Sigma$ (precisione normalizzata sull'osservazione)

$M_t \longleftrightarrow J^T \Sigma_o J + \Sigma$ (matrice di precondizionamento completa)

Il precondizionatore attenzionale M_t del Capitolo II è la matrice di Fisher della distribuzione congiunta dell'active inference; la maschera di attenzione A_t è la normalizzazione della componente diagonale della matrice di precisione sull'osservazione. Questa corrispondenza non è un'analogia: è un'identità strutturale. La geometria riemanniana del Capitolo II e la variational inference di Friston sono due linguaggi per lo stesso oggetto.

VI.3 La gerarchia delle precisioni come torre bayesiana con struttura attenzionale

Il framework gerarchico di Friston estende l'active inference a più livelli: lo stato al livello l è generato da una distribuzione condizionale $P(h^l \mid h^{l+1})$, e gli errori di predizione si propagano bottom-up mentre le predizioni si propagano top-down. Ogni livello ha la propria matrice di precisione Σ^l_o che pondera gli errori di predizione provenienti dal livello inferiore.

Questa struttura gerarchica è la realizzazione biologica della torre bayesiana gerarchica della terza monografia con struttura attenzionale. La corrispondenza è:

livello l della gerarchia di Friston $\longleftrightarrow$ livello π^{l-1} della torre bayesiana
matrice di precisione Σ^l_o $\longleftrightarrow$ maschera di attenzione A_t al livello l
ottimizzazione di Σ^l_o $\longleftrightarrow$ meta-learning della maschera attenzionale

Il punto cruciale che il saggio sull'energia libera non sviluppa nella direzione di questa monografia è che la *ottimizzazione* delle matrici di precisione — il processo con cui il sistema impara quali segnali pesare di più a ciascun livello della gerarchia — è un'istanza del meta-learning della maschera di attenzione: è il sistema che apprende quale sia la propria maschera ottimale per ogni livello della gerarchia predittiva.

Questo meta-learning è soggetto ai teoremi di impossibilità della terza monografia. Apprendere la matrice di precisione ottimale al livello l richiede un meta-modello della propria struttura attenzionale al livello l+1 — che richiede un meta-meta-modello al livello l+2 — e la torre delle precisioni ha esattamente la struttura del limite proiettivo della torre bayesiana, con il teorema di Chaitin che bopund la complessità descrittiva delle matrici di precisione apprendibili e il teorema di Löb che limita la capacità del sistema di auto-validare la propria struttura di precisione attraverso la propria struttura di precisione.

VI.4 Il residuo specifico: la sensibilità della precisione e la struttura dell'ansia

Il contributo specifico di questo capitolo al di là del saggio esistente è la formalizzazione della *sensibilità della precisione* — la dipendenza delle matrici di precisione dallo stato del sistema — come operatore che connette la dinamica emotiva con la struttura attenzionale.

In active inference, l'ansia è modellata come alta incertezza sul valore ottimale della precisione: il sistema non sa quale segnale pesare di più, il che produce oscillazioni nella matrice Σ_o e quindi nell'aggiornamento delle credenze. Formalmente, si definisce la *varianza della precisione* come:

$$\mathrm{Var}_\Pi(h) = \mathbb{E}_{\pi^1}[(\Sigma_o(h,h) - \mathbb{E}_{\pi^1}[\Sigma_o(h,h)])^2]$$

l'incertezza di secondo ordine sulla precisione dell'osservazione nella direzione dell'ipotesi h. La $\mathrm{Var}_\Pi(h)$ alta corrisponde a un sistema che non sa quanto pesare il segnale proveniente dalla direzione h — la struttura formale dell'ipervigilanza ansiosa: attenzione ampliata ma mal calibrata, che processa molti segnali senza poter determinare quali meritino aggiornamento pieno.

La connessione con la torre bayesiana è che $\mathrm{Var}_\Pi(h)$ è la varianza del meta-prior π^1 sulla maschera attenzionale nella direzione h: è esattamente l'incertezza di secondo ordine discussa nel paragrafo III.5 della terza monografia come "epistemic gap". Il sistema ansioso ha un epistemic gap positivo nella struttura della propria attenzione: sa di non sapere quanto prestare attenzione a certi segnali, e questa meta-incertezza si manifesta come oscillazione nell'aggiornamento delle credenze.

Capitolo VII — Flusso gradiente di Wasserstein e dinamica continua delle credenze

VII.1 Perché la divergenza KL non basta

Il framework del Capitolo II usa la divergenza KL come misura della distanza tra distribuzioni di credenze e il mirror descent come algoritmo di discesa del gradiente nello spazio P. Questa scelta è naturale per l'aggiornamento bayesiano — che è esattamente la proiezione KL sull'iperpiano della verosimiglianza — ma produce una metrica che non cattura un aspetto importante della dinamica delle credenze: il *costo di trasporto* della massa di probabilità da un'ipotesi a un'altra.

La divergenza KL misura la differenza "statistica" tra due distribuzioni: quanto sarebbe sorpreso un agente che usa σ nell'osservare campioni da π. Non misura quanto è "costoso" trasformare π in σ nel senso fisico: quanta massa di probabilità deve essere spostata, e di quanto, per passare da un'ipotesi all'altra. Per le credenze cognitive — distribuzioni concentrate su ipotesi con struttura metrica — il costo di trasporto è spesso la misura più naturale: aggiornare la credenza che h_1 sia vero verso la credenza che h_2 sia vero ha un costo che dipende dalla distanza tra h_1 e h_2 nello spazio delle ipotesi.

La distanza di Wasserstein W_2 cattura precisamente questo costo. Per due distribuzioni $\pi, \sigma \in P(H)$ su uno spazio metrico (H, d), la distanza di Wasserstein-2 è:

$$W_2(\pi, \sigma)^2 = \min_{\gamma \in \Gamma(\pi,\sigma)} \int_{H\times H} d(h, h')^2 \, d\gamma(h, h')$$

dove $\Gamma(\pi, \sigma)$ è l'insieme dei coupling di π e σ — le distribuzioni congiunta su $H \times H$ con marginali π e σ rispettivamente. W_2 è il minimo costo di trasporto della massa da π a σ, con costo quadratico nella distanza.

VII.2 La formula di Benamou-Brenier e l'energia cinetica del flusso di credenze

La formula di Benamou-Brenier (1998) fornisce una rappresentazione dinamica della distanza di Wasserstein-2 come problema di controllo ottimale:

$$W_2(\pi_0, \pi_1)^2 = \min_{ (\pi_t, v_t) : \partial_t \pi_t + \nabla\cdot(\pi_t v_t) = 0 } \int_0^1 \int_H \|v_t(h)\|^2 \, d\pi_t(h) \, dt$$

Il minimo è preso su tutte le traiettorie $(\pi_t)_{t \in [0,1]}$ che collegano π_0 a π_1 attraverso l'equazione di continuità $\partial_t \pi_t + \nabla\cdot(\pi_t v_t) = 0$, dove v_t è il campo di velocità del flusso. La distanza W_2 è la radice quadrata del minimo dell'energia cinetica integrata lungo la traiettoria.

L'interpretazione cognitiva è precisa. Il flusso di credenze (π_t) è la traiettoria delle credenze dell'agente nel tempo. L'equazione di continuità è la legge di conservazione della massa: la massa totale di probabilità è conservata durante il flusso. L'energia cinetica $\int\|v_t(h)\|^2 \, d\pi_t$ è il costo cognitivo del trasporto delle credenze: spostare credenza dalla regione h alla regione h' richiede una quantità di lavoro proporzionale a $\|h - h'\|^2$. La formula di Benamou-Brenier dice che W_2 è la distanza minima percorribile nel senso del minimo lavoro cognitivo.

Per il flusso di credenze sotto attenzione, l'equazione di continuità si modifica includendo il termine di sorgente dovuto all'aggiornamento attenzionato:

$$\partial_t \pi_t + \nabla \cdot (\pi_t v_t) = -\eta \cdot A_t \cdot (\log P(O_t \mid \cdot) - \mathbb{E}_{\pi_t}[\log P(O_t \mid \cdot)]) \cdot \pi_t$$

dove il lato destro è il termine di sorgente/assorbimento indotto dall'aggiornamento bayesiano con maschera A_t. Questo è il sistema che descrive la dinamica continua del flusso di credenze sotto osservazione continua con attenzione selettiva.

VII.3 Il JKO scheme: l'aggiornamento bayesiano come flusso gradiente di Wasserstein

Il JKO scheme (Jordan, Kinderlehrer, Otto, 1998) è il risultato che connette la distanza di Wasserstein con le equazioni di flusso gradiente. Il teorema fondamentale afferma:

Teorema JKO: L'equazione di diffusione (e più in generale le equazioni paraboliche della forma $\partial_t \pi = \nabla \cdot (\pi \nabla \delta F/\delta \pi)$) può essere generata come schema implicito di Eulero nello spazio di Wasserstein:

$$\pi_{t+\delta t} = \operatorname{argmin}_{\pi} \{ F(\pi) + W_2(\pi, \pi_t)^2/(2\delta t) \}$$

Il minimizzatore è la distribuzione che minimizza la somma del funzionale F e del costo di trasporto da π_t — è la distribuzione che fa il passo di discesa del gradiente di F nello spazio riemanniano (P, W_2).

Per il funzionale $F(\pi) = \mathbb{E}_\pi[-\log P(o|h)] + D_{KL}(\pi \| \pi_{prior})$ — la variational free energy dell'aggiornamento bayesiano — il JKO scheme produce:

$$\pi_{t+\delta t} = \operatorname{argmin}_{\pi} \{ \mathbb{E}_\pi[-\log P(o|h)] + D_{KL}(\pi \| \pi_{prior}) + W_2(\pi, \pi_t)^2/(2\delta t) \}$$

Questo è un aggiornamento bayesiano "lazy" nel senso di Wasserstein: aggiorna le credenze verso l'ottimo bayesiano ma penalizza gli spostamenti troppo grandi di massa di probabilità. La penalizzazione W_2 è il costo cognitivo del trasporto delle credenze — la resistenza del sistema a cambiare le proprie credenze proporzionale alla distanza metrica tra le ipotesi.

VII.4 Il flusso gradiente attenzionato nello spazio di Wasserstein

Il framework JKO si generalizza naturalmente all'aggiornamento attenzionato. Il flusso gradiente attenzionato nello spazio di Wasserstein è:

$$\pi_{t+\delta t}^A = \operatorname{argmin}_{\pi} \{ F^A(\pi, A_t) + W_2(\pi, \pi_t)^2/(2\delta t) \}$$

dove il funzionale attenzionato è:

$$F^A(\pi, A) = \int_H A(h) \cdot (-\log P(o|h)) \cdot \pi(h)\, dh + D_{KL}(\pi \| \pi_{prior})$$

il funzionale di costo dell'aggiornamento con la maschera A che pesa il termine di verosimiglianza.

Le condizioni di ottimalità del primo ordine per questo problema sono:

$$\log \pi(h) + A(h) \cdot \log P(o|h) + \log \pi_{prior}(h) + \varphi(h) = 0$$

dove φ è il potenziale di Kantorovich — la funzione duale del problema di trasporto ottimale — che emerge come moltiplicatore della penalizzazione W_2. La soluzione è:

$$\pi_{t+\delta t}^A(h) \propto \pi_{prior}(h) \cdot P(o|h)^{A(h)} \cdot \exp(-\varphi^*(h)/\delta t)$$

dove *$\varphi^(h)$ è la trasformata di Legendre-Fenchel del potenziale di Kantorovich. Il termine $\exp(-\varphi^/\delta t)$* è la "regolarizzazione spaziale" del trasporto: distribuisce la massa verso le ipotesi spazialmente vicine a quelle già credute, producendo un aggiornamento più morbido della semplice proiezione KL.

VII.5 La metrica di Wasserstein e la struttura dell'identità cognitiva stabile

Il framework di Wasserstein illumina la struttura della stabilità dell'identità cognitiva introdotta nel Capitolo V attraverso una lente geometrica diversa da quella KL. Nella metrica W_2, la distanza tra due distribuzioni di credenze è il costo di trasporto della massa — il costo "fisico" di spostare le credenze. Un sistema con identità cognitiva stabile ha un punto fisso μ^* che è stabile rispetto alla metrica W_2: piccole perturbazioni producono restoring forces che riportano il sistema verso μ^*.

La *curvatura di Ricci sintetica* nello spazio di Wasserstein — la condizione di curvatura $CD(K, \infty)$ di Lott-Villani-Sturm — determina la velocità con cui il flusso gradiente converge alla misura invariante:

$$W_2(\pi_t, \mu^*)^2 \leq e^{-2Kt} \cdot W_2(\pi_0, \mu^*)^2$$

Per $K > 0$ (curvatura positiva), la convergenza è esponenziale nel tempo con tasso K. La curvatura K è la misura formale della "stabilità dell'identità" nella metrica di Wasserstein: sistemi con alta curvatura (K grande) hanno identità più stabili che convergono più rapidamente al proprio attrattore dopo perturbazioni.

Il valore K dipende dalla struttura del funzionale F e dalla geometria dello spazio delle ipotesi H. Per funzionali F con forte componente di prior (alto $D_{KL}(\pi \| \pi_{prior})$),

K è proporzionale alla forza del prior — identità con prior forti sono più stabili. Per funzionali con debole componente di prior, K può essere vicino a zero o negativo: il sistema è neutralmente stabile o instabile rispetto alla metrica di Wasserstein.

VII.6 Confronto tra metrica KL e metrica W_2: quando usare ciascuna

Le due metriche su P — la divergenza KL e la distanza W_2 — catturano aspetti diversi della dinamica delle credenze e sono appropriate per diversi tipi di domande.

La divergenza KL è appropriata quando la domanda riguarda il *contenuto informazionale*: quanta sorpresa produce un campione da π per un agente che usa σ come riferimento? È la metrica naturale per l'aggiornamento bayesiano (proiezione sulla classe di verosimiglianza), per la calibrazione (distanza dal vero prior), e per il calcolo delle perdite informazionali dell'attenzione selettiva. Il mirror descent è il metodo ottimale nella metrica KL.

La distanza W_2 è appropriata quando la domanda riguarda il *costo di trasporto*: quanto lavoro cognitivo è necessario per spostare le credenze da π a σ? È la metrica naturale per misurare la resistenza dell'identità cognitiva (costo di allontanamento dal punto fisso), per la dinamica della persuasione (quanto "costoso" è cambiare opinione), e per la regolarizzazione spaziale dell'aggiornamento (flusso gradiente JKO). Il flusso gradiente di Wasserstein è il metodo ottimale nella metrica W_2.

Le due metriche producono traiettorie diverse per lo stesso problema di aggiornamento. La traiettoria KL-ottimale segue la geodetica e-piatta di Amari: aggiorna le statistiche sufficienti direttamente, indipendentemente dalla distanza metrica tra le ipotesi. La traiettoria W_2-ottimale segue la geodetica di Wasserstein: distribuisce la massa in modo da minimizzare il costo di trasporto, producendo aggiornamenti "morbidi" che si diffondono verso le ipotesi vicine a quelle già credute.

Il sistema cognitivo reale opera in un regime intermedio: l'aggiornamento bayesiano è KL-ottimale sul contenuto informazionale ma la rete neurale che lo implementa introduce una regolarizzazione W_2-type attraverso la continuità della mappa di attivazione del substrato subpropositivo. La geometria della varietà cognitiva $M \subset \mathbb{R}^n$ della seconda monografia è la struttura che media tra le due metriche: le strutture continue di M producono la regolarizzazione spaziale che avvicina la traiettoria delle credenze a quella W_2-ottimale.

Capitolo VIII — Gap spettrale, tempi di mescolamento e viscosità dell'identità cognitiva

VIII.1 La struttura spettrale del semigruppe dell'attenzione

Il semigruppe dell'attenzione $\{\Phi_t = \exp(\int_0^t \mathcal{A}_s\, ds) : t \geq 0\}$ introdotto nel paragrafo I.6 è un semigruppe di contrazioni su $L^2(P, \mu^*)$ — lo spazio delle funzioni di quadrato integrabile rispetto alla misura invariante μ^*. La struttura spettrale del generatore $\mathcal{A}$ determina la dinamica di decadimento delle perturbazioni attorno a μ^*.

Per sistemi con misura invariante μ^* che soddisfa la condizione di Poincaré — l'esistenza di una costante $\lambda_1 > 0$ tale che:

$$\mathrm{Var}_{\mu^*}(f) \leq (1/\lambda_1) \cdot \mathcal{E}(f, f) \quad \text{per ogni } f \in L^2(\mu^*)$$

dove $\mathcal{E}(f, f) = -\int f \cdot \mathcal{A}[f]\, d\mu^*$ è la forma di Dirichlet associata al generatore $\mathcal{A}$ — il semigruppe soddisfa:

$$\|\Phi_t(\pi) - \mu^*\|^2_{L^2(\mu^*)} \leq e^{-2\lambda_1 t} \cdot \|\pi - \mu^*\|^2_{L^2(\mu^*)}$$

Il gap spettrale λ_1 è il più piccolo autovalore non nullo di $-\mathcal{A}$: la distanza tra lo zero (autovalore della misura invariante) e il successivo autovalore del generatore. Un gap spettrale grande implica convergenza rapida alla misura invariante; un gap piccolo implica convergenza lenta.

VIII.2 Il calcolo del gap spettrale per maschere di attenzione a struttura prodotto

Per sistemi in cui la maschera di attenzione ha struttura prodotto — $A_t(h) = A_1(h_1) \cdot A_2(h_2) \cdot \ldots$ per ipotesi $h = (h_1, h_2, \ldots)$ fattorizzate — il generatore $\mathcal{A}$ si fattorizza in somma di generatori indipendenti sulle componenti:

$$\mathcal{A} = \mathcal{A}_1 \otimes I \otimes \ldots + I \otimes \mathcal{A}_2 \otimes \ldots + \ldots$$

Per semigruppii prodotto, gli autovalori del generatore complessivo sono le somme degli autovalori dei generatori componente. Il gap spettrale del semigruppe prodotto è:

$$\lambda_1^{tot} = \min_k \lambda_1^k$$

il minimo dei gap spettrali componenti. Il componente con il gap spettrale più piccolo determina il tempo di mescolamento globale: la "bottleneck" dello spazio delle credenze.

Per sistemi con struttura di grafo — dove le ipotesi sono nodi di un grafo $G = (H, E)$ e le transizioni sono permesse solo lungo gli archi — il gap spettrale è legato all'espansione del grafo. La disuguaglianza di Cheeger fornisce un bound:

$$\lambda_1 \geq h(G)^2 / 2$$

dove $h(G) = \min_{S \subset H} |\partial S| / \min(|S|, |H \setminus S|)$ è la costante isoperimetrica di Cheeger — la proporzione minima di massa che deve attraversare il confine tra due sottoinsiemi. Per grafi con buona espansione ($h(G)$ grande), il gap spettrale è grande e le credenze mescolano rapidamente. Per grafi con struttura a cluster ben separati ($h(G)$ piccolo), il gap è piccolo e il sistema rimane intrappolato in un cluster per lungo tempo.

VIII.3 Il tempo di mescolamento e la permanenza nei bacini di attrazione

Il tempo di mescolamento è definito come:

$$\tau_{mix}(\varepsilon) = \inf \{ t \geq 0 : \| \Phi_t(\pi) - \mu^* \|_{TV} \leq \varepsilon \text{ per ogni } \pi \in P \}$$

il primo tempo in cui tutte le distribuzioni iniziali si trovano entro ε dalla misura invariante in variazione totale. Il legame con il gap spettrale è:

$$\tau_{mix}(\varepsilon) \leq (1/\lambda_1) \cdot \log(1/\varepsilon + \|\pi_0 - \mu^*\|/\varepsilon)$$

e, con la disuguaglianza di Cheeger inversa:

$$\tau_{mix}(1/4) \geq 1/(2\lambda_1)$$

Il tempo di mescolamento è inversamente proporzionale al gap spettrale: sistemi con piccolo gap spettrale hanno tempi di mescolamento lunghi — restano intrappolati lontano dalla misura invariante per lungo tempo.

Per sistemi con struttura multimodale — misura invariante μ^* con due o più picchi separati da barriere — il gap spettrale è piccolo (proporzionale all'altezza della barriera): il sistema può restare intrappolato in una regione del paesaggio per tempi esponenzialmente lunghi. Questo è il meccanismo formale della *polarizzazione stabile*: una volta che la distribuzione delle credenze di un agente è concentrata in un bacino, il tempo di mescolamento verso la misura globale è esponenzialmente grande nell'altezza della barriera — il sistema non "esplora" spontaneamente credenze alternative.

VIII.4 Il gap spettrale come misura della viscosità dell'identità

Si definisce la *viscosità dell'identità cognitiva* come:

$$\zeta(A_t) = 1/\lambda_1(\mathcal{A}_t)$$

il reciproco del gap spettrale del generatore attenzionale. La viscosità misura quanto è "lento" il sistema a mescolare le proprie credenze sotto la dinamica dell'auto-attenzione: sistemi con alta viscosità restano in un bacino per lungo tempo; sistemi con bassa viscosità esplorano lo spazio delle credenze rapidamente.

La viscosità dipende dalla maschera di attenzione A_t attraverso la forma di Dirichlet $\mathcal{E}(f,f) = \int A_t(h) \cdot \|\nabla_h f(h)\|^2 \, d\mu^*(h)$. Una maschera A_t che sopprime le ipotesi vicino alle frontiere tra bacini ($A_t(h) \approx 0$ per h vicino alle separatrici) produce una forma di Dirichlet con small support sulle transizioni inter-bacino — il che riduce il gap spettrale e aumenta la viscosità. La conferma del bias è il caso estremo: $A_t^{confirm} \propto \pi_t$ sopprime esattamente le ipotesi nelle regioni a bassa credenza, che sono le regioni di frontiera tra bacini, producendo viscosità massima — il sistema si intrappola nel proprio bacino di attrazione.

Il meccanismo della terapia cognitivo-comportamentale può essere formalizzato come riduzione della viscosità: la terapia espone deliberatamente il sistema a stimoli nelle regioni di frontiera tra bacini (esposizione graduale, ristrutturazione cognitiva) — ipotesi che normalmente ricevono bassa attenzione perché lontane dal punto fisso corrente. Questo aumenta A_t nelle regioni di frontiera, aumenta la forma di Dirichlet sulle transizioni inter-bacino, riduce la viscosità λ_1^{-1}, e permette al sistema di esplorare bacini alternativi in tempi ragionevoli.

VIII.5 Il legame tra gap spettrale, prior di Solomonoff e profondità metacognitiva

Si stabilisce ora la connessione tra la struttura spettrale del semigruppe attenzionale e la terza monografia. Il gap spettrale λ_1 del semigruppe dell'attenzione è legato alla complessità algoritmica della misura invariante $KC(\mu^*)$:

Proposizione: Per sistemi con misura invariante μ^* a complessità $KC(\mu^*) = K$, il gap spettrale è bounded superiormente da:

$$\lambda_1 \leq C \cdot 2^{-K/d}$$

dove d è la dimensionalità intrinseca dello spazio delle ipotesi e C è una costante dipendente dalla struttura del sistema. Sistemi con misura invariante ad alta complessità hanno gap spettrale piccolo e alta viscosità.

La dimostrazione usa il bound di Cheeger inverso $h(G) \leq O(1/\sqrt{K})$: una misura invariante ad alta complessità algoritmica richiede un grafo di transizione con molte componenti debolmente connesse — la complessità della struttura di credenze si manifesta come bassa espansione del grafo e quindi piccolo gap di Cheeger.

Il corollario metacognitivo è: sistemi con identità cognitiva complessa (alta $KC(\mu^*)$) hanno alta viscosità — cambiano identità lentamente. Sistemi con identità semplice (bassa $KC(\mu^*)$) hanno bassa viscosità — cambiano identità rapidamente. La stabilità dell'identità è inversamente proporzionale alla sua semplicità algoritmica: le identità

robuste sono quelle complesse, non quelle semplici. Questo è un risultato controintuitivo che il framework formale rende preciso: la complessità cognitiva è una forma di protezione contro la manipolazione epistemica.

Capitolo IX — Fokker-Planck per le credenze collettive e Gibbard-Satterthwaite per i mercati dell'attenzione

IX.1 Il limite di campo medio e l'equazione di McKean-Vlasov

Il modello a N agenti del Capitolo IV descrive la dinamica delle credenze individuali sotto un campo attenzionale condiviso. Per $N \to \infty$, questo sistema ha un limite di campo medio: invece di tracciare le N distribuzioni individuali $\{\pi_i^t\}$, si considera la distribuzione empirica:

$$\varrho_N^t = (1/N)\, \Sigma_{i=1}^N \delta_{\pi_i^t}$$

la distribuzione su P che assegna massa 1/N a ogni credenza individuale. Per $N \to \infty$, ϱ_N^t converge debolmente a una misura di probabilità $\varrho^t \in P(P)$ — la distribuzione su P delle credenze nella popolazione — che evolve secondo una PDE nello spazio delle misure.

L'equazione che governa ϱ^t è l'equazione di McKean-Vlasov, una PDE con struttura di flusso gradiente nello spazio di Wasserstein su P:

$$\partial_t \varrho^t = \nabla_P \cdot (\varrho^t \cdot \nabla_P F[\varrho^t])$$

dove $F[\varrho] = \int_P F^{MF}(\pi, \varrho)\, d\varrho(\pi)$ è il funzionale di campo medio:

$$F^{MF}(\pi, \varrho) = \mathbb{E}_{\pi}[-\log P(O \mid h)] + D_{KL}(\pi \,\|\, \pi_{prior}) + \int_P V(\pi, \pi')\, d\varrho(\pi')$$

Il primo termine è il costo di verosimiglianza (aggiornamento bayesiano); il secondo è il costo rispetto al prior; il terzo è il termine di interazione — $V(\pi, \pi')$ è il potenziale di interazione tra due agenti con credenze π e π', che cattura l'influenza sociale tra agenti (imitazione, contrasto, dibattito).

IX.2 L'equazione di Fokker-Planck nello spazio delle distribuzioni

La versione stocastica dell'equazione di McKean-Vlasov include un termine di diffusione che modella il rumore nell'aggiornamento delle credenze individuali:

$$d\pi_i\hat{}t = -\nabla_P F[\varrho\hat{}t](\pi_i\hat{}t)\, dt + \sqrt{(2\beta^{-1})}\, dW_i\hat{}t$$

dove $W_i\hat{}t$ è un moto Browniano nello spazio P (il processo di Wiener nella metrica di Wasserstein) e β^{-1} è la temperatura — l'ampiezza del rumore stocastico nell'aggiornamento. L'equazione di Fokker-Planck per la densità $\varrho\hat{}t$ associata a questo sistema è:

$$\partial_t\, \varrho\hat{}t = \nabla_P \cdot (\varrho\hat{}t \cdot \nabla_P F[\varrho\hat{}t]) + \beta^{-1}\, \Delta_P\, \varrho\hat{}t$$

dove Δ_P è il laplaciano nello spazio P nella metrica di Wasserstein. La prima parte è il termine deterministico di discesa del gradiente; la seconda è la diffusione stocastica che impedisce la convergenza a distribuzioni degeneri.

Questa equazione è la PDE fondamentale della dinamica collettiva delle credenze: descrive come la distribuzione $\varrho\hat{}t$ delle credenze nella popolazione evolve nel tempo sotto la combinazione di aggiornamento bayesiano, interazione sociale e rumore stocastico. Le sue soluzioni stazionarie — le misure invarianti ϱ^* — soddisfano:

$$\nabla_P \cdot (\varrho^* \cdot \nabla_P F[\varrho^*]) + \beta^{-1}\, \Delta_P\, \varrho^* = 0$$

Le soluzioni di questa equazione sono le distribuzioni di Gibbs della forma:

$$\varrho^*(\pi) \propto \exp(-\beta \cdot F[\varrho^*](\pi))$$

distribuzioni di Boltzmann sul funzionale di campo medio: la probabilità di trovare un agente con credenza π è proporzionale a exp(-β · costo della credenza π dati gli altri). Per $\beta \to \infty$ (temperatura zero, no rumore), la distribuzione si concentra sui minimi del funzionale F — i punti fissi deterministici del campo medio. Per β finito, la distribuzione è "smussata" attorno ai minimi: c'è una probabilità non nulla di trovare agenti lontani dai minimi.

IX.3 La struttura delle soluzioni stazionarie: pluralismo e omologazione

Le soluzioni stazionarie dell'equazione di McKean-Vlasov determinano le strutture di credenza stabili della collettività. La struttura delle soluzioni dipende dalla forma del potenziale di interazione $V(\pi, \pi')$ e dalla temperatura β^{-1}.

Per potenziali di allineamento $V(\pi, \pi') = -\alpha \cdot \mathrm{sim}(\pi, \pi')$ (gli agenti sono attratti dagli agenti con credenze simili, come nelle camere eco digitali), il funzionale $F[\varrho]$ ha minimi che dipendono da α e β:

Per α piccolo (interazione sociale debole) e β grande (bassa temperatura): un unico minimo — la distribuzione bayesiana ottimale. Tutti gli agenti convergono alle stesse credenze accurate.

Per α grande (forte interazione di allineamento) e β grande: minimi multipli — distribuzione bimodale ϱ^* con due picchi separati. La collettività si polarizza in due cluster con credenze distanti. Ogni cluster è internamente allineato e lontano dall'altro: polarizzazione stabile.

Per α intermedio e β piccolo (alta temperatura): un minimo diffuso. La collettività mantiene alta diversità epistemica ma senza struttura: le credenze sono distribuite quasi uniformemente.

La transizione di fase — il passaggio da un unico minimo a minimi multipli al variare di α — è formalmente una *transizione di fase di McKean-Vlasov*, analoga alla transizione di fase nel modello di Curie-Weiss del ferromagnetismo. La temperatura critica:

$$\beta_c = 1 / (\alpha \cdot \chi_0)$$

dove χ_0 è la suscettibilità del sistema singolo (la varianza delle credenze di un singolo agente), determina il punto al di sotto del quale la polarizzazione emerge spontaneamente. Per $\beta > \beta_c$ (bassa temperatura, agenti con aggiornamenti poco rumorosi) e $\alpha > \alpha_c$ (forte interazione sociale), la polarizzazione è l'unico equilibrio stabile.

IX.4 La connessione con i mercati dell'attenzione: campo attenzionale come parametro d'ordine

Il campo attenzionale collettivo A_t^C del Capitolo IV agisce, nell'equazione di Fokker-Planck, come un parametro d'ordine che modula sia il termine di discesa del gradiente sia il termine di diffusione. La dipendenza di $F[\varrho^t]$ dalla maschera collettiva è:

$$F^A[\varrho](\pi) = \int_H A^C(h) \cdot (-\log P(O|h))\, d\pi(h) + D_{KL}(\pi \,\|\, \pi_{\{prior\}}) + \int V(\pi, \pi')\, d\varrho(\pi')$$

Il campo attenzionale A^C, determinato dal meccanismo di mercato del Capitolo IV, modifica selettivamente il gradiente del funzionale: le ipotesi con alta attenzione collettiva $A^C(h) > 1$ sono aggiornate più rapidamente, quelle con bassa attenzione sono aggiornate lentamente. Il mercato dell'attenzione, nella dinamica di Fokker-Planck, è un operatore che deforma il funzionale F — e quindi l'insieme dei suoi minimi e la struttura delle soluzioni stazionarie.

Un produttore che controlla A^C può scegliere il campo attenzionale per spostare le soluzioni stazionarie dell'equazione di Fokker-Planck verso Π^{target}: la distribuzione di credenze collettiva desiderata. Questo è il problema di controllo del

Capitolo III ora formulato nella forma completa: trovare A^C tale che la soluzione stazionaria ϱ*(A^C) sia vicina a Π^{target} nella metrica di Wasserstein su P(P).

Il problema è ben posto e ha soluzione sotto condizioni di regolarità di F. Il gradiente del funzionale rispetto al campo attenzionale A^C:

$$\delta/\delta A^C \; W_2(\varrho^*(A^C), \; \Pi^{target}) \; = \; -\mathbb{E}_{\varrho^*}[\mathbb{E}_{\pi}[\log \; P(O|h) \; \cdot \; \nabla_{A^C} \; \varrho^*(A^C)]]$$

è il gradiente del costo di trasporto dalla distribuzione stazionaria al target rispetto al campo attenzionale. Un produttore con potere epistemico Π(E, T) alto — che può controllare A^C per un'ampia porzione della collettività — può seguire questo gradiente per spostare iterativamente la distribuzione collettiva delle credenze verso il proprio target.

IX.5 Il teorema di Gibbard-Satterthwaite per i meccanismi di attenzione

Il teorema di Arrow del Capitolo IV stabilisce che nessuna funzione di aggregazione delle maschere di attenzione soddisfa simultaneamente dominio illimitato, Pareto, indipendenza e non-dittatura. Il *teorema di Gibbard-Satterthwaite* completa questo risultato dal lato strategico: nessun meccanismo di aggregazione è simultaneamente strategy-proof e non-dittatoriale.

Un meccanismo di aggregazione dell'attenzione Agg è *strategy-proof* se per ogni agente i, la maschera collettiva risultante Agg({A^i, A^{-i}}) non può essere resa più favorevole all'agente i attraverso la dichiarazione di una maschera distorta Ã^i ≠ A^i:

$$\mathcal{U}_i(\text{Agg}(A^i, A^{-i})) \geq \mathcal{U}_i(\text{Agg}(\tilde{A}^i, A^{-i})) \quad \text{per ogni } \tilde{A}^i \text{ e } A^{-i}$$

dove $\mathcal{U}_i$ è la funzione di utilità dell'agente i rispetto alla maschera collettiva.

Teorema di Gibbard-Satterthwaite per l'attenzione: Per spazi di ipotesi con |H| ≥ 3 e almeno tre agenti, nessuna funzione di aggregazione Agg che mappi insiemi di maschere individuali in maschere collettive è simultaneamente:

(SP) Strategy-proof: nessun agente ha incentivo a distorcere la propria maschera dichiarata
(ND) Non-dittatoriale: nessun agente determina unilateralmente la maschera collettiva
(SUR) Sur-jective: ogni possibile maschera collettiva è raggiungibile da qualche profilo di input

La dimostrazione è una riduzione al teorema originale di Gibbard (1973) e Satterthwaite (1975). Si costruisce una corrispondenza tra il problema di aggregazione

delle maschere di attenzione e un problema di scelta sociale standard: ogni maschera A^i di un agente è un "tipo" che determina le sue preferenze sull'insieme delle maschere collettive possibili (l'agente preferisce le maschere collettive che si avvicinano alla propria). La struttura di questo problema soddisfa le ipotesi del teorema di Gibbard-Satterthwaite, che si applica direttamente.

Il corollario è che qualsiasi meccanismo di mercato dell'attenzione — algoritmo di raccomandazione, sistema editoriale, meccanismo di voto — non è strategy-proof: esiste sempre un profilo di maschere individuali per cui almeno un agente ha incentivo a dichiarare una maschera diversa da quella genuina. I mercati dell'attenzione sono vulnerabili alla manipolazione strategica da parte degli agenti più sofisticati — quelli che capiscono la struttura del meccanismo di aggregazione e possono scegliere la propria dichiarazione attenzionale per massimizzare la maschera collettiva a proprio favore.

IX.6 La cascata di manipolazione strategica

Il teorema di Gibbard-Satterthwaite produce una struttura dinamica di manipolazione strategica a cascata. Quando un agente i capisce che il meccanismo non è strategy-proof e inizia a dichiarare Ã^i ≠ A^i, gli altri agenti osservano il comportamento di i e apprendono che i sta manipolando. A questo punto hanno incentivo a manipolare a loro volta. Il risultato è un equilibrio di Nash del gioco di manipolazione strategica — un profilo di maschere dichiarate {Ã^i*} che è un equilibrio del gioco in cui ogni agente sceglie la propria maschera dichiarata per massimizzare il proprio payoff dato il comportamento degli altri.

Nell'equilibrio di Nash del gioco di manipolazione, la maschera collettiva Agg({Ã^i*}) non riflette in generale le maschere genuine {A^i}: è la maschera prodotta dalla manipolazione strategica reciproca, che può divergere significativamente dall'aggregato genuino. La struttura dell'equilibrio dipende dalla distribuzione del potere epistemico: gli agenti con alto potere epistemico (quelli che riescono a influenzare la maschera collettiva più efficacemente) convergono a un equilibrio in cui la maschera collettiva si avvicina alla loro maschera genuina, a scapito degli agenti con basso potere epistemico.

Questo produce un risultato di concentrazione dinamica del potere epistemico: anche partendo da una distribuzione uniforme del potere, la manipolazione strategica induce una concentrazione verso gli agenti più sofisticati (con potere epistemico strutturalmente maggiore), che a loro volta acquisiscono ancora più potere attraverso la propria capacità di influenzare la maschera collettiva. Il mercato dell'attenzione è un meccanismo che amplifica le diseguaglianze del potere epistemico nel tempo.

Appendici tecniche

Appendice A — Il mirror descent nello spazio di Bregman: derivazioni

A.1 La famiglia esponenziale come geodetiche e-piatte

Nello spazio delle distribuzioni di probabilità su H, la famiglia esponenziale:

$$P_\theta(h) = \exp(\theta \cdot T(h) - A(\theta))$$

dove $T(h)$ è la statistica sufficiente e $A(\theta) = \log \int \exp(\theta \cdot T(h))\, d\mu(h)$ è la funzione log-partizione, forma una sotto-varietà e-piatta. Le coordinate naturali θ e le coordinate medie $\eta = \mathbb{E}_{P_\theta}[T(h)]$ sono duali rispetto alla metrica di Fisher: la metrica ha la forma:

$$g_{ij}(\theta) = \partial^2 A / \partial\theta_i \partial\theta = \mathrm{Cov}_{P_\theta}[T_i(h), T(h)]$$

La matrice di Fisher coincide con la matrice di covarianza della statistica sufficiente — un risultato della geometria dell'informazione che ha la seguente interpretazione metacognitiva: l'incertezza sulle statistiche sufficienti (la loro varianza sotto P_θ) è identica alla curvatura dello spazio delle credenze in θ.

A.2 Il passo di mirror descent come proiezione KL

Il passo di mirror descent con potenziale negentropico $\Psi(\pi) = \Sigma_h\, \pi(h) \log \pi(h)$ è:

$$\pi_{t+1} = \mathrm{argmin}_{\pi} \{ \eta \langle \nabla \mathscr{F}(\pi_t), \pi \rangle + D_{KL}(\pi \,\|\, \pi_t) \}$$

Le condizioni di ottimalità del primo ordine sono:

$$\eta\, \nabla \mathscr{F}(\pi_t)(h) + \log \pi_{t+1}(h) - \log \pi_t(h) + \lambda = 0$$

dove λ è il moltiplicatore del vincolo di normalizzazione. Risolvendo per π_{t+1}:

$$\pi_{t+1}(h) = \pi_t(h) \cdot \exp(-\eta\, \nabla \mathscr{F}(\pi_t)(h) - \lambda)$$
$$\propto \pi_t(h) \cdot \exp(-\eta\, \nabla \mathscr{F}(\pi_t)(h))$$

Per $\nabla \mathscr{F}(\pi_t)(h) = -\log P(o \mid h)$ (gradiente della log-verosimiglianza negativa):

$$\pi_{t+1}(h) \propto \pi_t(h) \cdot \exp(\eta \log P(o \mid h)) = \pi_t(h) \cdot P(o \mid h)^{\eta}$$

che è l'aggiornamento bayesiano generalizzato con tasso η.

A.3 La convergenza del mirror descent con gradiente attenzionato

Per il mirror descent con gradiente attenzionato $\nabla_P\hat{}A\ \mathcal{F}$, la convergenza al minimizzatore π^* della funzione obiettivo è garantita sotto condizioni di Lipschitz sul gradiente. Il bound sulla convergenza è:

$$\mathbb{E}[\mathcal{F}(\bar{\pi}_T) - \mathcal{F}(\pi^*)] \leq D_KL(\pi^* \,\|\, \pi_0) \,/\, (\eta\, T) + \eta\, L^2/2$$

dove $\bar{\pi}_T = (1/T)\ \Sigma_\{t=1\}\hat{}T\ \pi_t$ è la media temporale, L è la costante di Lipschitz del gradiente, e il rate è $O(1/\sqrt{T})$ per $\eta = O(1/\sqrt{T})$. La perdita di calibrazione dovuta all'attenzione selettiva rallenta la convergenza di un fattore $(1/\|A\|_\infty)$: maschere con $A(h) < 1$ riducono il tasso effettivo di apprendimento.

Appendice B — La geometria dell'informazione di Amari: connessioni duali

B.1 Le due connessioni affini

Nello spazio delle distribuzioni di probabilità P, la geometria dell'informazione di Amari (1985) identifica due connessioni affini duale:

La connessione e-affine (esponenziale) $\nabla\hat{}\{(1)\}$ ha simboli di Christoffel:

$$\Gamma_i\square\square\hat{}\{(1)\} = \mathbb{E}_P[\partial_i\partial\ \log P \cdot \partial\ \log P]$$

La connessione m-affine (mistura) $\nabla\hat{}\{(-1)\}$ ha simboli di Christoffel:

$$\Gamma_i\square\square\hat{}\{(-1)\} = -\mathbb{E}_P[\partial_i\partial\ \log P \cdot \partial\ \log P] + \mathbb{E}_P[\partial_i \log P \cdot \partial\ \log P \cdot \partial\ \log P]$$

Le due connessioni sono duali rispetto alla metrica di Fisher: $\nabla\hat{}\{(1)\} + \nabla\hat{}\{(-1)\} = 2\ \nabla\hat{}\{(0)\}$ (la connessione di Levi-Civita della metrica di Fisher). Questa dualità produce la *struttura statistica duale* dello spazio P.

B.2 La non-commutatività delle proiezioni m-piatte

Due proiezioni m-piatte consecutive su iperpiani L_1 e L_2 producono risultati diversi a seconda dell'ordine:

$$\Pi\hat{}m_\{L_2\}(\Pi\hat{}m_\{L_1\}(\pi)) \neq \Pi\hat{}m_\{L_1\}(\Pi\hat{}m_\{L_2\}(\pi))$$

a meno che $L_1 \perp L_2$ nella metrica di Fisher. La commutatore delle due proiezioni è:

$$[\Pi\hat{}m_\{L_1\}, \Pi\hat{}m_\{L_2\}](\pi) \approx R\hat{}\{(-1)\}(\partial L_1, \partial L_2) \cdot \varepsilon^2 + O(\varepsilon^3)$$

dove R^{(-1)} è il tensore di curvatura della connessione m-piatta valutato nelle direzioni tangenti agli iperpiani, e ε è l'ampiezza della variazione. Per spazi e-piatti (famiglie esponenziali), R^{(-1)} è non nullo: le proiezioni non commutano anche nelle sottofamiglie esponenziali, il che è la fonte formale del primacy e del recency effect dell'aggiornamento attenzionato.

Appendice C — Teoria dei giochi del mercato dell'attenzione

C.1 Il gioco del mercato dell'attenzione

Si formalizza il mercato dell'attenzione come un gioco in forma normale. I giocatori sono i K produttori di segnali {P_k}_{k=1}^K. Lo spazio delle strategie di ogni produttore è lo spazio dei segnali {s_k ∈ $\mathscr{S}$_k} che può generare, con costo c_k(s_k). La funzione di payoff del produttore k è:

u_k(s_k, {s_{k'}}_{k'≠k}) = w_k({s_{k'}}_{k'}) · R_k(s_k)

dove w_k è la quota di attenzione catturata dal produttore k (dipende dall'intero profilo di segnali) e R_k(s_k) è il ricavo unitario per unità di attenzione. La quota di attenzione è determinata da una funzione di salienza:

w_k = σ(s_k, {s_{k'}}_{k'}) / Σ_{k'} σ(s_{k'}, {s_{l}}_{l})

una normalizzazione softmax sulle salienze. L'equilibrio di Nash è il profilo {s_k*}_{k} tale che nessun produttore può aumentare il proprio payoff modificando unilateralmente il proprio segnale.

C.2 Il bias sistemico dell'equilibrio di Nash verso la salienza

Sotto la funzione di salienza σ(s, π_i) = exp(β · emotività(s)) + γ · allineamento(s, π_i) (combinazione di emotività e allineamento con le credenze esistenti), l'equilibrio di Nash produce segnali con le seguenti proprietà:

emotività(s_k*) → max (massimizzazione dell'impatto emotivo)
informatività(s_k*) → irrilevante (la calibrazione non entra nel payoff)
polarizzazione(s_k*) → max (massimizzazione della risposta degli estremi)

L'equilibrio di Nash del mercato dell'attenzione con payoff-salienza produce segnali massimalmente emotivi e polarizzanti, indipendentemente dal loro contenuto informazionale. Questo è il teorema di inefficienza epistemica del mercato dell'attenzione: il meccanismo di mercato non seleziona per la calibrazione delle credenze ma per la cattura dell'attenzione, e le due proprietà sono sistematicamente

disallineate sotto qualsiasi funzione di salienza che privilegia emotività e allineamento rispetto all'informatività.

Appendice D — L'auto-attenzione nei transformer: connessione con il framework formale

D.1 Il meccanismo di self-attention nei modelli di linguaggio

Il meccanismo di self-attention dei modelli Transformer (Vaswani et al., 2017) ha la struttura formale:

$$\text{Attention}(Q, K, V) = \text{softmax}(QK^T / \sqrt{d_k}) \cdot V$$

dove Q (query), K (key), V (value) sono trasformazioni lineari dell'input. Il termine $\text{softmax}(QK^T / \sqrt{d_k})$ è una matrice di pesi di attenzione che determina quali posizioni nell'input "guardano" quali altre posizioni. Nella nomenclatura del framework: la matrice softmax è una versione discreta e parametrizzata della maschera di attenzione A_t, dove le ipotesi H sono le posizioni nella sequenza e le osservazioni sono i token.

D.2 La connessione con il gradiente naturale

Il meccanismo di self-attention produce pesi proporzionali a:

$$A_\{ij\} \propto \exp(q_i \cdot k / \sqrt{d_k})$$

che è una distribuzione di Boltzmann sulla compatibilità query-key. Questo corrisponde, nel framework del gradiente naturale, a una maschera di attenzione che approssima la matrice di Fisher: le posizioni con alta compatibilità query-key corrispondono alle coppie di ipotesi tra cui il sistema ha alta informazione mutua. Il meccanismo di self-attention è un'implementazione approssimata del gradiente naturale attenzionato in cui la matrice di precondizionamento è approssimata dalla struttura query-key.

D.3 I limiti dell'auto-attenzione dei transformer rispetto all'auto-attenzione metacognitiva

Il meccanismo di self-attention dei Transformer è un'auto-attenzione di primo ordine: il sistema determina la propria maschera in funzione dell'input corrente. Non è auto-attenzione di secondo ordine nel senso del Capitolo V: il sistema non forma credenze sulla propria maschera attenzionale, non aggiorna quelle credenze in risposta all'osservazione del proprio comportamento, non ha un meta-modello della propria struttura attenzionale. Il punto fisso dell'auto-attenzione transformer è determinato

dai pesi appresi (Q, K, V parametrizzati) e non dalla dinamica di auto-osservazione descritta nel Capitolo V.

Questa distinzione è rilevante per la quarta monografia nella sua connessione con la seconda: i modelli di linguaggio operano senza substrato subpropositivo (seconda monografia) e senza meta-modello della propria struttura attenzionale (quarta monografia). La loro auto-attenzione è sintattica — determina quali token influenzano quali nella sequenza — ma non metacognitiva nel senso precisato: il sistema non ha credenze sulla propria maschera di attenzione che vengono aggiornate attraverso l'osservazione del proprio comportamento.

Appendice E — Il flusso gradiente di Wasserstein: derivazioni

E.1 Il teorema di Benamou-Brenier: derivazione

Il teorema di Benamou-Brenier afferma che $W_2(\pi_0, \pi_1)^2$ è il minimo dell'energia cinetica lungo le traiettorie che soddisfano l'equazione di continuità.

Si prova prima che $W_2(\pi_0, \pi_1)^2$ è un lower bound per l'energia cinetica. Per ogni traiettoria (π_t, v_t) con $\partial_t \pi_t + \nabla \cdot (\pi_t v_t) = 0$ che collega π_0 a π_1, il coupling γ_t definito da:

$$\int \varphi(h,h')\, d\gamma_t = \int_0^t \int_H \varphi(h,\, h + \int_0^t v_s(X_s^h)\, ds)\, d\pi_0(h)\, dt$$

è un coupling di π_0 e π_t. Per il teorema di Cauchy-Schwarz:

$$W_2(\pi_0, \pi_t)^2 \le \int_{H\times H} \|h - h'\|^2\, d\gamma_t \le t \cdot \int_0^t \int_H \|v_s(h)\|^2\, d\pi_s(h)\, ds$$

Prendendo $t = 1$ si ottiene $W_2(\pi_0, \pi_1)^2 \le \int_0^1 \int_H \|v_t\|^2\, d\pi_t\, dt$.

Per l'uguaglianza: il coupling ottimale γ^* di π_0 e π_1 — il trasporto ottimale — definisce il campo di velocità ottimale $v_t^*(h) = (T^*(h) - h)$ per le geodetiche rettilinee, dove T^* è la mappa di trasporto ottimale. Lungo questa traiettoria, l'energia cinetica è esattamente $W_2(\pi_0, \pi_1)^2$.

E.2 Il JKO scheme: la variante attenzionata

Per il funzionale attenzionato $F^A(\pi, A) = \int H\, A(h)(-\log P(o|h))\pi(h)dh + D_{KL}(\pi \| \pi_{prior})$, il JKO scheme produce:

$$\pi_{t+\delta t}^A = \operatorname{argmin}_{\pi} \{ F^A(\pi, A_t) + W_2(\pi, \pi_t)^2/(2\delta t) \}$$

Le condizioni di ottimalità richiedono l'uguaglianza:

$$\delta F^A/\delta\pi(h) + \varphi(h) = C$$

dove φ è il potenziale di Kantorovich e C è una costante di normalizzazione. Calcolando $\delta F^A/\delta\pi$:

$$A_t(h) \cdot (-\log P(o|h)) + \log \pi(h)/\pi_{prior}(h) + 1 + \varphi(h) = C$$

quindi:

$$\log \pi_{t+\delta t}^A(h) = \log \pi_{prior}(h) + A_t(h) \cdot \log P(o|h) - \varphi(h) + C'$$

$$\pi_{t+\delta t}^A(h) \propto \pi_{prior}(h) \cdot P(o|h)^{A_t(h)} \cdot \exp(-\varphi(h))$$

Il termine $\exp(-\varphi(h))$ è la regolarizzazione di trasporto: il potenziale di Kantorovich φ distribuisce la massa verso le ipotesi vicine a quelle già credute, producendo la "morbidezza" caratteristica del flusso W_2.

E.3 La condizione di curvatura CD(K, ∞) e la convergenza

La condizione di curvatura-dimensione CD(K, ∞) di Lott-Villani-Sturm per lo spazio metrico di misura (H, d, μ) afferma che per ogni coppia $\pi_0, \pi_1 \in P(H)$ e ogni geodetiche di Wasserstein $(\pi_t)_{t \in [0,1]}$:

$$\mathrm{Ent}(\pi_t \,\|\, \mu) \leq (1-t)\,\mathrm{Ent}(\pi_0 \,\|\, \mu) + t\,\mathrm{Ent}(\pi_1 \,\|\, \mu) - (K/2)\,t(1-t)\,W_2(\pi_0, \pi_1)^2$$

dove $\mathrm{Ent}(\pi \,\|\, \mu) = \int \log(d\pi/d\mu)\, d\pi$ è l'entropia relativa. Per $K > 0$, questa condizione implica la contrazione esponenziale del flusso gradiente:

$$W_2(\Phi_t(\pi), \mu^*)^2 \leq e^{-2Kt}\, W_2(\pi, \mu^*)^2$$

La costante K è la "curvatura di Ricci sintetica" dello spazio: misura quanto rapidamente le geodetiche si avvicinano. Per $K = 0$, la contrazione è sub-esponenziale; per $K < 0$, la divergenza è esponenziale.

Appendice F — Equazione di Fokker-Planck per le credenze collettive: struttura e soluzioni

F.1 Il limite di campo medio: derivazione rigorosa

Si derivano le equazioni di McKean-Vlasov dal modello a N agenti. Ogni agente i evolve secondo:

$$d\pi_i^t = b(\pi_i^t, \varrho_N^t)\, dt + \sqrt{(2\beta^{-1})}\, dW_i^t$$

dove $b(\pi, \varrho) = -\nabla P$ *Fϱ è il drift e W_i^t sono moti Browniani indipendenti. La distribuzione empirica $\varrho_N^t = (1/N)\, \Sigma_i\, \delta\{\pi_i^t\}$* soddisfa, per $N \to \infty$:

Teorema di propagazione del caos (McKean, 1966): Sotto condizioni di Lipschitz su b, per ogni k fisso e $N \to \infty$, la distribuzione marginale di k agenti converge al prodotto $(\varrho^t)^{\otimes k}$, dove ϱ^t è la soluzione dell'equazione di McKean-Vlasov:

$$\partial_t \varrho^t = -\nabla_P \cdot (\varrho^t \cdot b(\cdot, \varrho^t)) + \beta^{-1}\, \Delta_P\, \varrho^t$$

La propagazione del caos afferma che, nel limite $N \to \infty$, gli agenti diventano (asintoticamente) indipendenti e identicamente distribuiti secondo ϱ^t: la distribuzione collettiva è descritta da una singola PDE nonlineare anziché da un sistema di N equazioni accoppiate.

F.2 L'analisi di stabilità lineare e la transizione di fase

Si analizza la stabilità della soluzione omogenea ϱ^*_0 — la distribuzione stazionaria in assenza di interazione ($V = 0$) — rispetto a perturbazioni che attivano l'interazione.

Per potenziale $V(\pi, \pi') = -\alpha \cdot \langle \pi, \pi' \rangle$ (inner product delle credenze come misura di allineamento), la linearizzazione dell'equazione di McKean-Vlasov attorno a ϱ^*_0 produce:

$$\partial_t\, \delta\varrho = L_0[\delta\varrho] + \alpha \cdot T[\delta\varrho]$$

dove L_0 è il generatore dell'equazione di Fokker-Planck senza interazione e $T\delta\varrho = \int \langle \pi, \pi' \rangle\, \varrho_0(\pi')\, d\pi'$ *è l'operatore di interazione linearizzato. La stabilità di* ϱ_0 richiede che il più grande autovalore di $L_0 + \alpha T$ sia negativo.

Il più grande autovalore di $L_0 + \alpha T$ è:

$$\lambda_{max} = -\lambda_1(L_0) + \alpha \cdot \sup_{f:\, \|f\|=1} \langle Tf, f \rangle_{L^2(\varrho^*_0)}$$

La transizione di fase si verifica quando $\lambda_{max} = 0$:

$$\alpha_c = \lambda_1(L_0) / \|T\|_{op}$$

Per $\alpha > \alpha_c$, la soluzione omogenea è instabile e il sistema converge a una soluzione eterogenea — una distribuzione polarizzata ϱ^* con più picchi.

F.3 L'equazione di Fokker-Planck con campo attenzionale esterno

Con campo attenzionale A^C(h) variabile nel tempo controllato da un produttore esterno, l'equazione di Fokker-Planck diventa:

$$\partial_t \rho^t = -\nabla_P \cdot (\rho^t \cdot \nabla_P F^{A^C(t)}[\rho^t]) + \beta^{-1} \Delta_P \rho^t$$

Questa è una PDE non-autonoma: il funzionale F dipende dal parametro esterno A^C(t). Il problema di controllo ottimale del campo attenzionale è:

$$\min_{A^C(t)} \int_0^T W_2(\rho^t, \Pi^{target})^2 \, dt$$ soggetto all'equazione di Fokker-Planck

Le condizioni di ottimalità del primo ordine producono un sistema di equazioni forward-backward: la PDE di Fokker-Planck (forward) per ρ^t e l'equazione di Hamilton-Jacobi-Bellman (backward) per la funzione di valore $V^t(\rho)$. La risoluzione numerica di questo sistema è computazionalmente costosa ma ben definita in linea di principio.

Appendice G — Il teorema di Gibbard-Satterthwaite per l'attenzione: derivazione

G.1 Formulazione del teorema come problema di scelta sociale

Si riduce il problema dell'aggregazione delle maschere di attenzione a un problema di scelta sociale standard. Ogni agente i ha tipo $\theta_i = A^i \in H_A$ (la propria maschera genuina). L'insieme delle alternative è $X = H_A$ (l'insieme delle possibili maschere collettive). La funzione di scelta sociale $f: H_A^N \to H_A$ è la funzione di aggregazione Agg.

Le preferenze dell'agente i sull'insieme X sono indotte dalla propria maschera genuina A^i: l'agente i preferisce la maschera collettiva A a A' se l'aggiornamento sotto A è più favorevole per i rispetto all'aggiornamento sotto A'. Si assume che queste preferenze soddisfino la proprietà di ordinamento: per ogni agente i, la funzione di preferenza indotta è un ordine totale su X.

G.2 Applicazione del teorema di Gibbard-Satterthwaite

Il teorema originale afferma: per $|X| \geq 3$, ogni funzione di scelta sociale f che sia:

(i) Surjective: ogni alternativa in X è scelta per qualche profilo di tipi
(ii) Strategy-proof: dichiarare il proprio tipo genuino è una strategia dominante

è dittatoriale: esiste un agente i tale che $f(\theta) = \theta_i$ per ogni profilo θ.

L'applicazione al problema delle maschere: per $|H_A| \geq 3$ (almeno tre maschere distinte possibili) e $N \geq 3$, qualsiasi funzione di aggregazione $Agg: H_A^N \to H_A$

che sia surjective e strategy-proof è dittatoriale. Poiché la dittatura è incompatibile con la non-dittatura (ND), ne segue che nessun meccanismo di aggregazione soddisfa simultaneamente surjettività, strategy-proofness e non-dittatura.

G.3 La struttura dell'equilibrio di Nash della manipolazione strategica

Nel gioco in cui ogni agente sceglie la propria maschera dichiarata per massimizzare il proprio payoff, la struttura dell'equilibrio di Nash dipende dalla funzione di payoff e dalla regola di aggregazione. Per aggregazione lineare $A\hat{}C = \Sigma_i w_i \cdot A\hat{}i$, l'equilibrio di Nash è:

$\tilde{A}\hat{}i^* = \text{argmax}_\{\tilde{A}\hat{}i\}\ \mathcal{U}_i(\Sigma_i w_i \tilde{A}\hat{}i + \Sigma_\{j\neq i\}\ w\ \ A\hat{}j^*)$

Sotto la funzione di utilità $\mathcal{U}_i(A\hat{}C) = -\|A\hat{}C - A\hat{}i\|^2$ (ogni agente vuole minimizzare la distanza della maschera collettiva dalla propria), la condizione di ottimalità è:

$w_i \cdot \partial A\hat{}C/\partial\tilde{A}\hat{}i = -2w_i\ (A\hat{}C - A\hat{}i)$

Per aggregazione lineare $\partial A\hat{}C/\partial\tilde{A}\hat{}i = w_i$, la condizione diventa:

$w_i^2 = -2w_i\ (A\hat{}C - A\hat{}i) \rightarrow\ A\hat{}C = A\hat{}i + w_i/2 \cdot I_perp$

dove I_perp è la componente di identità ortogonale alla maschera genuina A^i. Nell'equilibrio di Nash, ogni agente dichiara una maschera che è la propria genuina più una perturbazione nella direzione dell'identità, scalata per il proprio peso w_i. Gli agenti con peso alto (potere epistemico alto) distorcono di più la propria dichiarazione; quelli con peso basso distorcono di meno. L'equilibrio produce una maschera collettiva che è distorta nella direzione degli agenti ad alto peso — amplificando la concentrazione del potere epistemico.

Bibliografia

Ottimizzazione stocastica e mirror descent

Nemirovski, A., Yudin, D. (1983). *Problem Complexity and Method Efficiency in Optimization*. Wiley, New York. Il testo fondativo dell'ottimizzazione in spazi di Bregman — il framework del Capitolo II.

Beck, A., Teboulle, M. (2003). Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3), 167-175. La

formulazione moderna del mirror descent e la sua convergenza — base dell'Appendice A.

Shalev-Shwartz, S. (2012). Online learning and online convex optimization. *Foundations and Trends in Machine Learning*, 4(2), 107-194. Il testo che connette il mirror descent con l'aggiornamento bayesiano online.

Geometria dell'informazione

Amari, S. (1985). *Differential-Geometrical Methods in Statistics*. Springer, Berlino. Il testo fondativo della geometria dell'informazione — la struttura riemanniana dello spazio delle distribuzioni sviluppata nel Capitolo II.

Amari, S., Nagaoka, H. (2000). *Methods of Information Geometry*. American Mathematical Society / Oxford University Press. Il testo di riferimento per le connessioni duali e la struttura duale dello spazio delle distribuzioni — base dell'Appendice B.

Amari, S. (2016). *Information Geometry and Its Applications*. Springer, Tokyo. Il testo più recente che connette la geometria dell'informazione con i sistemi di apprendimento — la connessione con il gradiente naturale del Capitolo II.

Gradiente naturale e apprendimento

Amari, S. (1998). Natural gradient works efficiently in learning. *Neural Computation*, 10(2), 251-276. L'articolo che introduce il gradiente naturale come metodo di apprendimento efficiente — il fondamento del paragrafo II.1.

Martens, J. (2020). New insights and perspectives on the natural gradient method. *Journal of Machine Learning Research*, 21(146), 1-76. La revisione moderna del gradiente naturale con connessioni con il mirror descent e l'ottimizzazione bayesiana.

Economia dell'attenzione

Kahneman, D. (1973). *Attention and Effort*. Prentice-Hall, Englewood Cliffs NJ. Il testo fondativo della psicologia cognitiva dell'attenzione come risorsa limitata — il framework standard che questa monografia estende e critica.

Simon, H.A. (1971). Designing organizations for an information-rich world. In: Greenberger, M. (Ed.), *Computers, Communication, and the Public Interest*, 37-72, Johns Hopkins University Press. Il testo in cui Simon introduce il concetto di economia dell'attenzione — la scarsità dell'attenzione come risorsa strategica in un mondo di abbondanza informativa.

Wu, T. (2016). *The Attention Merchants: The Epic Scramble to Get Inside Our Heads*. Knopf, New York. La storia del mercato dell'attenzione moderna — la base empirica dell'analisi formale del Capitolo III.

Polarizzazione e dinamiche collettive delle credenze

Pariser, E. (2011). *The Filter Bubble: What the Internet Is Hiding from You*. Penguin, New York. La documentazione del meccanismo di amplificazione della salienza nei media digitali — il fenomeno che il teorema di inefficienza epistemica dell'Appendice C formalizza.

Sunstein, C.R. (2017). *#Republic: Divided Democracy in the Age of Social Media*. Princeton University Press. L'analisi delle conseguenze sistemiche della polarizzazione epistemica sulla democrazia deliberativa.

Bakshy, E., Messing, S., Adamic, L.A. (2015). Exposure to ideologically diverse news and opinion on Facebook. *Science*, 348(6239), 1130-1132. Lo studio empirico sulla struttura dell'esposizione informativa nelle piattaforme social — dati per la validazione del modello di polarizzazione del Capitolo IV.

Manipolazione e filosofia della persuasione

Marlin, R. (2002). *Propaganda and the Ethics of Persuasion*. Broadview Press, Peterborough. L'analisi etica della propaganda che la distinzione formale del paragrafo III.4 formalizza.

Cialdini, R.B. (1984). *Influence: The Psychology of Persuasion*. William Morrow, New York. Il testo che cataloga i meccanismi psicologici della persuasione — le istanze empiriche dei tre operatori di influenza epistemica del Capitolo III.

Self-attention e transformer

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30. L'articolo originale dell'architettura Transformer — la base dell'Appendice D.

Tsai, Y.H.H., Bai, S., Liang, P.P., Kolter, J.Z., Morency, L.P., Salakhutdinov, R. (2019). Multimodal transformer for unaligned multimodal language sequences. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 6558-6569. Estensione del meccanismo di self-attention a input multimodali — rilevante per la connessione con le architetture multimodali della seconda monografia.

Connessioni con le monografie precedenti del ciclo

Lewis, D. (1969). *Convention: A Philosophical Study*. Harvard University Press, Cambridge MA. Il testo che formalizza i giochi di coordinazione e le convenzioni comunicative — la base della prima monografia su cui il framework attenzionale si innesta per il problema del coordinamento dell'attenzione.

Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, 11(2), 127-138. Il principio di energia libera come framework unificante per la cognizione bayesiana — la connessione tra aggiornamento bayesiano, attenzione, e substrato subpropositivo che attraversa le quattro monografie.

Hohwy, J. (2013). *The Predictive Mind*. Oxford University Press. La teoria del cervello predittivo come sistema che minimizza l'errore di predizione bayesiano — il framework in cui il gradiente attenzionato del Capitolo I trova la propria implementazione biologica.

Teoria dell'attenzione e neuroscienze computazionali

Corbetta, M., Shulman, G.L. (2002). Control of goal-directed and stimulus-driven attention in the brain. *Nature Reviews Neuroscience*, 3(3), 201-215. La base neurobiologica dei due sistemi attenzionali — top-down (guidato dagli obiettivi) e bottom-up (guidato dagli stimoli) — la cui interazione è formalizzata come ottimizzazione della maschera A_t nel Capitolo I.

Desimone, R., Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annual Review of Neuroscience*, 18(1), 193-222. Il modello biased competition dell'attenzione visiva — il modello biologico più vicino alla struttura formale della maschera bang-bang ottimale del Capitolo I.

Posner, M.I., Petersen, S.E. (1990). The attention system of the human brain. *Annual Review of Neuroscience*, 13(1), 25-42. Il testo fondativo del sistema attenzionale umano con la distinzione tra orientamento, detezione e allerta — le componenti funzionali che il framework formale unifica come variazioni della maschera A_t.

Trasporto ottimale e distanza di Wasserstein

Villani, C. (2003). *Topics in Optimal Transportation*. American Mathematical Society, Providence RI. Il testo fondativo del trasporto ottimale moderno — la base della distanza W_2 e della formula di Benamou-Brenier del Capitolo VII.

Villani, C. (2008). *Optimal Transport: Old and New*. Springer, Berlino. Il testo comprensivo sul trasporto ottimale, inclusa la teoria di Lott-Villani-Sturm sulla curvatura sintetica — la base della condizione $CD(K,\infty)$ dell'Appendice E.

Benamou, J.D., Brenier, Y. (2000). A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem. *Numerische Mathematik*, 84(3), 375-393. L'articolo che introduce la formulazione dinamica della distanza W_2 — il teorema centrale del Capitolo VII.

Jordan, R., Kinderlehrer, D., Otto, F. (1998). The variational formulation of the Fokker-Planck equation. *SIAM Journal on Mathematical Analysis*, 29(1), 1-17. L'articolo originale del JKO scheme — il risultato che connette il flusso gradiente di Wasserstein con la dinamica delle distribuzioni.

Santambrogio, F. (2015). *Optimal Transport for Applied Mathematicians*. Birkhäuser, Basel. Il testo applicativo sul trasporto ottimale con particolare attenzione al flusso gradiente di Wasserstein e alle sue applicazioni alle PDE.

Equazioni di McKean-Vlasov e campo medio

McKean, H.P. (1966). A class of Markov processes associated with nonlinear parabolic equations. *Proceedings of the National Academy of Sciences*, 56(6), 1907-1911. L'articolo fondativo della propagazione del caos e delle equazioni di McKean-Vlasov — la base teorica del Capitolo IX.

Sznitman, A.S. (1991). Topics in propagation of chaos. In: Hennequin, P.L. (Ed.), *Ecole d'Eté de Probabilités de Saint-Flour XIX*, Springer, 165-251. Il testo di riferimento per la teoria rigorosa della propagazione del caos — la base matematica del limite di campo medio.

Chizat, L., Bach, F. (2018). On the global convergence of gradient descent for over-parameterized models using optimal transport. *Advances in Neural Information Processing Systems*, 31. La connessione tra McKean-Vlasov e l'ottimizzazione di reti neurali sovra-parametrizzate — rilevante per la connessione con i sistemi cognitivi collettivi.

Gap spettrale e tempi di mescolamento

Levin, D.A., Peres, Y., Wilmer, E.L. (2009). *Markov Chains and Mixing Times*. American Mathematical Society, Providence RI. Il testo di riferimento per la teoria dei tempi di mescolamento — la base dei risultati del Capitolo VIII.

Diaconis, P., Stroock, D. (1991). Geometric bounds for eigenvalues of Markov chains. *Annals of Applied Probability*, 1(1), 36-61. Il risultato classico che connette la disuguaglianza di Cheeger con il gap spettrale — la base dell'Appendice E relativa al gap spettrale per grafi.

Bakry, D., Émery, M. (1985). Diffusions hypercontractives. In: *Séminaire de Probabilités XIX*, Lecture Notes in Mathematics, 1123, 177-206, Springer. Il testo fondativo della teoria Γ_2 per i processi di diffusione — la connessione tra curvatura e gap spettrale per le equazioni di Fokker-Planck.

Gibbard-Satterthwaite e meccanismi di incentivo

Gibbard, A. (1973). Manipulation of voting schemes: A general result. *Econometrica*, 41(4), 587-601. L'articolo originale del teorema di Gibbard — metà della coppia Gibbard-Satterthwaite del Capitolo IX.

Satterthwaite, M.A. (1975). Strategy-proofness and Arrow's conditions: Existence and correspondence theorems for voting procedures and social welfare functions. *Journal of Economic Theory*, 10(2), 187-217. L'articolo originale di Satterthwaite — l'altra metà della coppia.

Moulin, H. (1980). On strategy-proofness and single peakedness. *Public Choice*, 35(4), 437-455. Il risultato che identifica i meccanismi strategy-proof per preferenze con picco singolo — la classe di preferenze attenzionali per cui vale la strategy-proofness sotto restrizioni appropriate.

Precision weighting e active inference: aspetti collaterali

Feldman, H., Friston, K.J. (2010). Attention, uncertainty, and free-energy. *Frontiers in Human Neuroscience*, 4, 215. L'articolo che formalizza l'attenzione come ottimizzazione della precisione nell'active inference — la base della connessione del Capitolo VI.

Rao, R.P.N., Ballard, D.H. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, 2(1), 79-87. Il testo fondativo del predictive coding gerarchico — il framework biologico in cui la gerarchia delle precisioni del Capitolo VI è implementata.

Mathys, C.D., Lomakina, E.I., Daunizeau, J., Iglesias, S., Brodersen, K.H., Friston, K.J., Stephan, K.E. (2014). Uncertainty in perception and the hierarchical Gaussian filter. *Frontiers in Human Neuroscience*, 8, 825. Il filtro gaussiano gerarchico come implementazione computazionale della gerarchia delle precisioni — la connessione tra la struttura della torre bayesiana e la neurobiologia.